Project M9: Exoplanet detection using the transit method

Marko Raidlo, Raido Everest

Business understanding

Background

An exoplanet is defined as a planet which orbits a star outside our own solar system. Exoplanets are a field of research for astronomers and astrophysicists to gather more data on different planetary systems and make better theories on system formation and evolution. Also of great interest is the possibility of extraterrestrial life existing. Studying exoplanets could help answer the age-old question: are we alone in the universe?

The Kepler space telescope is a retired NASA probe launched in 2009 which had a goal of detecting earth-like exoplanets using the transit photometry method. The telescope with a 1,4 meter primary mirror would take measurements of the sky every 6 seconds and average the measurements out for a minute for short cadence targets (512 targets per campaign) and 30 minutes for long cadence targets (170 000 targets per campaign). Measurements would usually last for 3 - 4 months, until the telescope would have to rotate itself in order to be oriented away from the star. These measurements of the stars allow us to detect slight decreases in the brightness of the stars. The decreases are usually caused by a planet passing in front of the star and blocking out some of its light.

In the next decade, the Hubble Space Telescope's successor, the James Webb Space Telescope, is set to launch with 4 other space telescopes (CHEOPS, PLATO, ARIEL, WFIRST), all of which have primary or secondary goals of exoplanet detection and research. Also, future space telescopes are designed such as the enormous LUVIOR (Large UV Optical Infrared Surveyor) with a 15 meter primary mirror. All these missions will produce large amounts of data for astronomers to analyze. Training efficient models to find candidate exoplanets from the light curve data will improve our abilities to discover new exoplanets as well as find candidates for future direct imaging missions. There is definitely room for improvement with used classification models, as a human volunteer from Planet Hunters TESS managed to find a couple of exoplanets missed by professional astronomers by manually sorting through 17517 measurements of NASAs TESS satellite.

Business goals

The goal of this project is to find the most efficient machine learning or deep learning models for classification of exoplanet candidates based on light curve data from space telescopes. Ultimately the project benefits us individually, by learning how to handle time series data like this.

Business success criteria

Successful application of theories learned in the course, by creating accurate exoplanet classification models. Members of the project gather insight into data mining processes. This'll be graded by whichever grading committee from university ends up grading the project.

Assessing your situation

- **Inventory of resources**
    - 2 team members
    - Kepler time series data from Kaggle
    - Python library stack: matplotlib, seaborn, numpy, pandas, scikit-learn, possibly keras with tensorflow or PlaidML backend and other such freely available libraries.

- **Requirements, assumptions, and constraints**
  We need to complete this project by the preset presentation date of 19th December. We need to have completed at the very least our primary goal of detecting exoplanet presence based on time series data.

- **Risks and contingencies**
  Insufficient knowledge of processing time series. Can be remedied by online resources. Insufficient domain knowledge as well, but in this case, not too much information is required to be learned.

- **Terminology**
    - Transit - A phenomenon in which a celestial body passes between an observer and the observed object.
    - Transit photometry - A technique to detect transit in which we measure the light coming off from an object over a period of time.
- **Costs and benefits**
  Due to the project not being business-focused, this really isn't a factor.
    - **Costs:**
        - Basically, the electricity to run a laptop, which is negligible.
    - **Benefits:**
        - Possibly, future hireability of the persons working on the project, which has an expected value of perhaps at least a couple of cents higher than the electricity.

Data-mining goals

A model to detect stars with exoplanets based on existing time series data.

If we dip into NASA's data sources, perhaps a new dataset similar to the Kaggle one containing some extra information that may be useful.

Data-mining success criteria

Our primary model should do well enough, arbitrary goal .8 recall and precision, but of course we may go further.

Data understanding

Gathering data

For the purposes of finding if there is an exoplanet and to potentially find its radius, transit photometry is used. We require access to a dataset that includes a time series of observations of flux of the stars from a certain position. The data, of course, has to have been gathered long enough for the exoplanet to actually get between the probe and the star.

The data is readily available in several formats, available direct from NASA, which will be quite difficult to actually use, but also we can find time series data on Kaggle.

Ultimately, we'll see what we can do using the Kaggle data. The time series information from it is without a doubt the most important part of the project. Should need really strike, it might be the case that we'll gather information from the NASA Exoplanet Science Institute that includes also information about the star itself, however it will be necessary to figure out how to get only a certain subset of the data, due to there simply being so much of it.

The dataset available on Kaggle is quite simple to handle and will fit into the memory of a computer quite easily. NASA data is, however, largely published in a rather large amount of separate .fits files, the gathering and processing of which may be a rather tremendous undertaking. As such, we may limit ourselves to the data available on Kaggle.

Describing data

I shall henceforth be describing the dataset available through Kaggle. The data is laid out in quite a simple fashion - we have a column for whether our observed star has an exoplanet or not, and we have a time series containing measured flux from the star. As it stands, the dataset is ready to be worked with. It includes information for nearly six thousand stars. This data should be enough to fulfil our primary goal of detecting whether a star has an exoplanet orbiting it or not.

Exploring data

We do not have many variables in terms of what we are measuring, even if there's thousands of columns. We have the label column denoting whether the star has an exoplanet or not. The rest of the columns are there to house the time series.

The data itself usually plots something like a straight line, with the flux being a bit higher at times and lower at others. You could say without anything interesting going on, it's quite close to a standard distribution, should you rescale and average it all. Most of the time, the data stays only between +20 and -20.

Verifying data quality

The data possess quite a lot of spiking as well, including, at times, upwards. The data itself is denoised by NASA before even being made public by using algorithms to get rid of any potential artifacting by the probes themselves.

Planning

Main tasks:

- **Project preparations (2 x 4 h)**
  Complete homework 10, explore the data set, do research into different methods used by astronomers for this task, etc.
- **Cleaning up data / Data Augmentation (2 x 2 h)**
  The Kepler dataset is imbalance because of the small number of positive classification. Different algorithms (such as SMOTE) could be used to make more data with positive classification.
- **Feature engineering (Fourier transform or some other methods) (2 x 6 h)**
  Use different methods to transform time series data into different useful features using which, classification models can be made.
- **Training models (2 x 12 h)**
  Select and train machine learning and deep learning models. Apply parameter tuning and different methods to avoid model overfitting.
- **Interpretation of results (2 x 3 h)**
  Evaluate the success of different models.
- **Creating the poster (2 x 3 h)**
  Present the results of the project in the form of a poster.

Bonus tasks: (if there is time)

- Find a way to convert NASA .fits files into time series
- Apply models to the new data

Methods:

- Linear Regression
- Logistic Regression
- Support Vector Machine
- K Nearest Neighbours
- Decision trees
- Random forest
- Clustering
- Different time series analysis tools
- Neural networks