MLDI Assignment 1 - 2018

All assignment information, dates, submission portal and late policies are in the online resources. Students are encouraged to consult with each other on the questions, however all submitted work must reflect an individual student's contribution. You may attempt this assignment in any programming language, however Python is suggested (it will be necessary for Assignment 2). ALL PROGRAMS WRITTEN TO COMPLETE THIIS ASSIGNMENT MUST BE SUBMITTED ALSO.

QUESTION 1

Astronomers on the planet Proxima b have made an interesting discovery – the presence of an exoplanet around nearby star 'Sol'. An eccentric Proximian billionaire (Glaxorg the Opulent) plans to send a fleet of laser-powered spacecraft to visit this new planet - named Sol b - to (ostensibly) establish contact with any life-forms living there. But before launching the mission, Glaxorg the Opulent wants to check if the planet is at least capable of supporting life.

Proximian scientists believe that life is possible if liquid water exists on the planet's surface. This requires the planet to be in the so-called habitable-zone, the range of distances from the star where the temperature is suitable for liquid water to exist. The answer may lie in the discovery-data taken by Proximian telescopes…

Sol b was discovered using the radial-velocity method, wherein the Doppler-shift in spectral lines emitted from the host star is measured (via a spectrograph) as a function of time. If a planet is orbiting the star, it should slightly tug upon the star as it orbits, causing the star to periodically move towards and away from the observer. This changing line-of-sight velocity (aka radial velocity, RV) causes the Doppler-shift in spectral lines to take on a characteristic sinusoidal form.

a) Download the file proximian.csv from the Assignment site. Data consists of a series of measurements over time (given in seconds from some arbitrary start), with the Doppler Velocity given in m/sec. Plot the data with error bars. Don't forget to include labels!

Our goal is to fit the Proximian data. We will assume that the data can be described by a sine of the following form:

$$f(P, K, \phi) = K\sin(Pt + \emptyset)$$

And we will attempt to determine that values of P, K and ϕ that produce a sine function that best fits our Proximian data.

We can measure how good a sine function is by calculating the sum of the squares of the error which is given by the following formula:

$$\text{Error} = \sum_i^N (o_i - f_i)^2$$

Where N is the number of data points, o_i is the i^{th}observed data point and f_i is the i^{th}calculated data point.

b) Write a script that calculates the error for a given P, K and ϕ. Test your code on the following inputs:

	P	K	ϕ
1	10^7	1	0
2	10^6	0.01	0
3	5×10^7	2	π

c) Write a genetic algorithm (GA) to estimate values for P, K and ϕ using the error given above as your fitness score. You should give a brief statement describing how you are performing selection, crossover and mutation as well as any other features of your GA. You should also list all your hyperparameters, their corresponding value and give a brief description of each one.

Initiate your population within the following bounds:

$$0 < P < 10^8, \qquad 0 < K < 0.2, \qquad 0 < \phi < 2\pi$$

State the values of P, K and ϕ for the best individual you find seeding the random number generator at 0, 1 and 2. Plot the fitness of the best individual at each generation when you seed the random number generator at 0.

While GA's are good at exploring for regions of significance, in order to actually find the best fit with properly weighted errors, MCMC is often employed, leveraging information from the GA.

d) Write a program to perform a Monte-Carlo Markov Chain in order to estimate the parameters P, K and ϕ. Verify that both GA and MCMC find the same optimum by starting your chain at some (modest) distance from the best solution found by your GA (for example say about 2e7 in P and .04 in K). Produce a 2-D plot in your chosen pair of dimensions (e.g. P and K) on which you clearly identify the "burn in" of the chain.

e) Run the chain for 10^4 iterations. Make 3 plots showing the chain on a scatter plot (as a function of the two parameters). The plots should have the respective (X, Y) axis: (P, K), (P, ϕ) and (K, ϕ).

f) Plot the marginal likelihoods for P, K and ϕ. (Plot these as histograms with 50 bins, remember to discard the burn-in). What is your estimate of the most likely values for each parameter and give your error?

g) Make a plot of the raw data (with error bars), and overplot the sine function using your determined values of P, K and ϕ.

h) The Proximian astronomers have determined that Sol is a G type star, with a mass $M_{star} = 2 \times 10^{30}$ kg. Given its expected luminosity, they predict the habitable zone around this star to be between about 8×10^7 and 3×10^8 km from the star. According to Kepler's 3rd Law, a planet's distance to the star r is given by

$$r^3 = \frac{GM_{star}}{4\pi^2} P^2$$

where G is the universal gravitational constant. What is the orbital radius of Sol b? Is it in the habitable zone?

i) Despite Glaxorg the Opulent's stated scientific goals, it is suspected he actually wants to mine the planet for its delicious minerals.

The actual velocity of the star V_{star} is given by

$$K = V_{star} \sin i$$

where i is the orbital inclination. Assume the Sol system is seen edge-on (so i = 90 degrees). The velocity of an orbiting planet V_{pl} is given by

$$V_{pl} = \sqrt{GM_{star}/r}$$

the planet mass is then given by

$$M_{pl} = \frac{M_{star}V_{star}}{V_{pl}}$$

Assuming every part of Sol b is a desirable mineral, how much material does Glaxorg stand to obtain?

THERE ARE NO MORE QUESTIONS