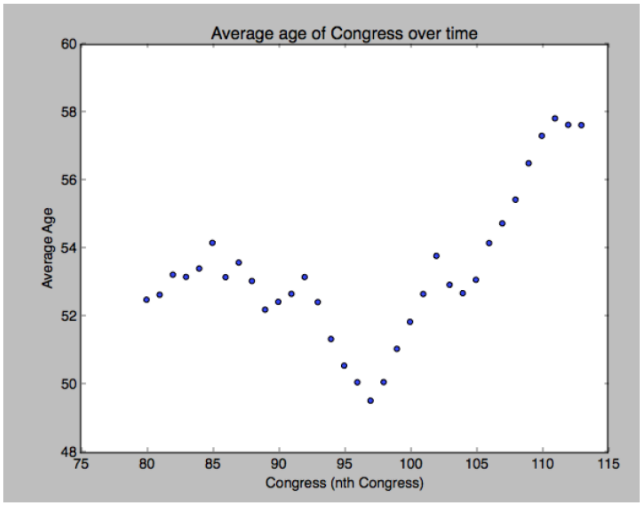
Jordan University of Science and Technology
CS763: Machine Learning/ Spring 2019
Dr. Malak Abdullah
Assignment 2
Due Date: 4-March-2019 (23:55:00)

Problem1:

The congress-ages.csv contains the data you will use for this problem. It has two columns. The first one is an integer that indicates the Congress number (This data had been taken from Harvard university). The second is the average age of that members of that Congress. The data would look like this:

```
congress, average_age
80,52.4959
81,52.6415
82,53.2328
83,53.1657
84,53.4142
85,54.1689
86,53.1581
87,53.5886
```

And you can see a plot of the data as:



X= nth Congress   y = Average age

Implement basis function regression with ordinary least squares with the above data. Some sample Python code is provided in linreg.py, which implements linear regression. Plot the data and regression lines for the simple linear case, and for each of the following sets of basis functions:

(a) $\phi_j(x) = x^j$ for $j = 1, \ldots, 6$

(b) $\phi_j(x) = x^j$ for $j = 1, \ldots, 4$

(c) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 6$

(d) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 10$

(e) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 22$

In addition to the plots, provide one or two sentences for each with numerical support, explaining whether you think it is fitting well, over fitting or under fitting. If it does not fit well, provide a sentence explaining why. A good fit should capture the most important trends in the data.

You should also submit the code

I will recommend that you read the following references:
https://newonlinecourses.science.psu.edu/stat501/node/324/
http://blog.robofied.com/polynomial-regression/
http://scipy-lectures.org/intro/scipy/auto_examples/plot_curve_fit.html

Extra for Geeks: https://fkorona.github.io/ATAI/2017_1/Lecture_notes/03-01_Linear_basis_function_models.pdf

Problem2:
Implement linear regression from scratch (Python code without using Scikit learn)

$$Y = \beta_0 + \beta_1 X$$

$$\beta_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$RMSE = \sqrt{\sum_{i=1}^{m} \frac{1}{m}(\hat{y}_i - y_i)^2}$$

1- Write a python function Coefficients that takes two vectors (X, Y) and returns B0 and B1
2- Write a function that calculates RMSE that takes (X, Y, B0, B1)
3- Test the above functions and compare them with Scikit learn linear regression with one variable. Use file2.csv for comparison and plotting