CSC110Y1-F Fall 2020 - Fundamentals of Computer Science 1
Course Project Proposal

Ching Chang Letian Cheng Arkaprava Choudhury Hanrui Fan

December 13, 2020

1. Part 1

An interactive report of the effect of Amazon rainforest's area on the local annual precipitation and CO_2 emission (with python prime shipping).

Ching Chang, Letian Cheng, Arkaprava Choudhury, Hanrui Fan

2. Part 2

Research Question: **To what extent do the changes in the Amazon rainforest's area affect the local annual precipitation and CO_2 emission?**

During our initial research on topics related to climate change, we found that the South American rainforest contributes 20% (Thomas, 2020) of the oxygen produced by photosynthesis on land, while the Amazon rainforest is responsible for 10% of the current greenhouse gas emissions (Melillo et al., y 20). This information was surprising to us since 20% of photosynthesis implies a lot of conversion from carbon dioxide, which is a greenhouse gas, to oxygen, whereas the 10% contribution to greenhouse gas seems to contradict that information. Due to this contradiction, we were curious about whether tree populations actually help control the climate. After some research, we learned that the effect of trees on climate change is more complex than we originally thought. There are many factors to consider, such as the carbon dioxide to oxygen conversion, the tendency to trap heat due to their dark color, reaction to form methane and ozone, and political movements revolving around tree plantations (Marshall, y 26). This led us into choosing an empirical approach —we wanted to directly observe the relationship between the change of tree population and climate change. We chose to focus our data on the Amazon rainforest not only because it is the largest rainforest on Earth (World Wildlife Fund., 2013), but also because there have been several pieces of evidence that show that the Amazon rainforest has been suffering from deforestation recently. For example, over $700,000km^2$ $(270,000mi^2)$ of Amazon rainforest had been lost since 1970, reducing its size to 80.7% of its original size, in 2018 (Butler, ry 4); There have been more than 40,000 fires in the rainforest in 2019 (Government of Brazil., er 4); and that forest exploitation in Amazon has risen for 14 consecutive months in June 2020 (Reuters, 2020). With these major pieces of evidence of deforestation correlating to the change in global and local climate, we believe that it is a relevant topic to contemporary society that should not be ignored.

3. # Part 3

itemlize

Part 4

TODO: INPUT DESCRIPTION.

We first create a function that parses the `html` element of the stats on the website as a string, and converts it to a `nested array` so that it's easier to work with. This will involve using a `for loop`, `if statements`, and an `accumulator` keeping track of the data parsed so far. Using this function, we will collect our data for deforestation in the Amazon rain-forest over the past few decades. With the data now converted into a form that we can easily manipulate, we shall focus on analysing the data using our own functions.

TO REVIEW: ANALYSIS DESCRIPTION.

For this project, we use smooth polynomial fitting to relate two of the variables in our `nested list`. Now, although there exist readily available functions that would do the same in the module `numpy` (SciPy community, e 29), we try implementing our own functions for the same, to test our learning from the course.

We split the mathematical algorithm for this problem using top-down design. The most important part of the data analysis section is the PolynomialRegression custom class that we have defined, which, when initializing a variable of this class, takes in a dictionary with the name of the independent variable as the key, and its corresponding value in the dictionary as the list of all its values, and also, a corresponding dictionary for the dependent variable, and also a specified degree for the polynomial, and the level of precision expected. The initializer then calls upon a function that we designed using the ordinary least squares method to create a list of coefficients for the polynomial function that has the least variance with respect to the input data. Note that this list is implemented in terms of ascending order of power, to favour the implementation of the other methods in this class that analyze how well this polynomial represents the data (i.e., a measure of the accuracy of this polynomial). We defined the method `__call__` to allow the polynomial to defined as a callable, i.e., if `poly` is a variable of type PolynomialRegression, we we can simply evaluate `poly` at a value `x` by calling `poly(x)`.

We first have the method `r_squared` to calculate the coefficient of determination for this polynomial. We also have `extreme_absolute_error` to return the maximum and minimum absolute difference between the polynomial's value and a particular value in the input data. Then, we have the additional methods `covariance_with_polynomial` and `correlation_of_data` as additional measure of accuracy.

The other part in this section of the project is defining our own functions to handle the matrices involved in such a computation. We have the functions `make_matrix` and `find_coefficients` which are perhaps the most important functions in this part. The former initializes the matrix X, and the latter solves the matrix equation $\vec{y} = X\vec{\beta} + \epsilon$ for β, where X represents the matrix of powers of x created from the input data, $\vec{y}$ represents the values of y in the input data, and $\vec{\beta}$ is the estimated coefficients vector, while $\vec{\epsilon}$ consists of the random errors at each of the points (i.e., the difference between the polynomial evaluated at the point and the actual y-value). Since we could not find any efficient way to create an algorithm to compute the inverses (we haven't really covered any such 'efficient' algorithm in either MAT223 or MAT240), we made use of `numpy`'s built-in function called `numpy.linalg.inv()` along with a type conversion method `tolist()` that converts objects of type array (a data class in `numpy`) to a nested list format.

Now, finally, we also added a method in the PolynomialRegression class to allow for plotting the graph of the polynomial and a scatter plot of the data, using the functions `numpy.linspace` `matplotlib.pyplot`, and specifying our own customizations to the graph.

TODO: OUTPUT DESCRIPTION.

In addition to the graph, we plan to create an interactive text-based report of our data, where the user inputs a value for the independent or dependent variable, and the program will provide the corresponding dependent or independent value, coefficient of determination, or the slope of the tangent at the point, depending on which one the user asks for. The output will be text-based, and will require string concatenation, and if statements to check whether to add trivial information to the report.

The input/output model will use while loops and input prompt to keep the program interactive. We also extrapolate the data to yield predictions about future data using the interactive i/o model. Finally, we use the extrapolated data to summarize the upcoming significant years where the dependent variable will reach a certain milestone.

TODO: MAIN DESCRIPTION.

Technical requirement:

The Annual total CO_2 emissions dataset we have is a CSV file. In order to parse it easily, We decide to use pandas for CSV file reading.

Pandas is a python library written for data manipulation and analysis. (Pandas Development Team., 2020) In particular, it offers data structures and operations for manipulating numerical tables and time series.

Specific, we will use the function `read_csv`". This function has many parameters available for us, but we will only use `filepath_or_buffer` parameter. It takes a valid string path, which is the path to our CSV file.

This function will return a class - Pandas DataFrame that contains the data of the CSV file. Pandas DataFrame is easier for calculating dataset thanks to its various built-in functions. This way, we can access each cell of the CSV file for our later research.

Part 5

Part 6

For the data analysis part of this project, we initially planned on using an approach that was quite similar to the calculus 'way' of constructing the polynomial regression (hereby shortened to simply 'polyreg') model. For instance, in the project proposal, we mentioned that we would first construct a trivial initial candidate polynomial (we were, of course, referring to the linear regression model from Assignment 1, when we said trivial, since it has the right 'shape' but nor the right degree), which would then use to find the sum of the squares of the perpendicular distance from each of the points to the polynomial. We had also mentioned that we would then minimize this sum by using the simplex algorithm.

However, we later found out that coding the simplex algorithm and modifying it to fit this task was not a managaeable feat, as there so many degrees of freedom for this problem that we simply could not write an efficient piece of code to work in all cases.

As such, we switched our approach to the more 'linear algebra way' of solving this task, by using matrices instead. We made good use of the Gauss-Markov theorem [**TODO: Add citations**, and applied it to the ordinary least squares method **TODO: Add citations** to arrive at a relatively straightforward, already well-established, and mathematically proven approach that was more guaranteed to offer a good estimate for the polynomial model.

Part 7

Notes for discussion of polyreg: TODO: Add note on whether it is appropriate to use polyreg for extrapolation.

Notes for discussion of output / interaction: TODO: Add discussion points

References

Butler, R. (2020, January 4). Calculating deforestation figures for the amazon. *Mongabay*. https://rainforests.mongabay.com/amazon/deforestation_calculations.html.

Government of Brazil. (2020, November 4). Queimadas. *INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS*. http://queimadas.dgi.inpe.br/queimadas/portal-static/situacao-atual/.

Marshall, M. (2020, May 26). Planting trees doesn't always help with climate change. *BBC*. https://www.bbc.com/future/article/20200521-planting-trees-doesnt-always-help-with-climate-change.

Melillo, J., McGuire, A., Kicklighter, D., Moore III, B., and Vörösmarty, C. (1993, May 20). Global climate change and terrestrial net primary production. *Nature*. 363:234–240. https://www.bbc.com/future/article/20200521-planting-trees-doesnt-always-help-with-climate-change.

Pandas Development Team. (2020). pandas documentation. https://pandas.pydata.org/docs/.

Reuters, T. (2020). Brazil amazon deforestation up in june, set for worst year in over a decade. *CBC*. https://www.cbc.ca/news/world/amazon-deforestation-up-june-1.5644730.

SciPy community (2020, June 29). numpy.polyfit. https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html.

Thomas, A. (2020). Biodiversity and the amazon rainforest. *Greenpeace*. https://www.greenpeace.org/usa/biodiversity-and-the-amazon-rainforest/.

World Wildlife Fund. (2013). Our world's largest rainforest: The amazon [youtube]. *Youtube*. https://www.youtube.com/watch?v=bYAZ3NWVgtc.