CS532 Final Project Activity
Climate Data Fitting and Local Warming Justification

Junda Chen, Haoruo Zhao, Doris Duan

December 14, 2018

1 Abstract

The project introduce a simplified model to justify whether global warming is truly an issue in the current society. Student will first intensify their knowledge about **Basis Matrix** – its construction and its application to the training process. Student will then construct different basis matrix for training, based on their knowledge of global warming, and apply Ridge Regression and LASSO to possibly find the dominant factors of global warming. Student will compare the two methods in their ability to select dominant factors of global warming, and then try to justify the authenticity of local warming by scientific soundness of the factor they choose. *Optionally, student will continue to apply their model to other areas around the globe and try to approach the solution to justify global warming (given sufficient time range).*

Student will be able to master the construction of Basis Matrix in similar real-world problem, and apply the method in compatible with simple machine learning model such as Ridge Regression and LASSO. Meanwhile, the activity also alarm the pervasive use of machine learning in data analysis – researchers shall never overuse machine learning as the silver bullet to naively derive scientific conclusion.

2 Background

!! Before you start: The code could fail due to incompatibility with old version. The matlab code is built upon version R2017a, and some of the functionality is not supported in R2016. We have tried hard to be compatible, but if you're running on these versions and encounter a problem, please contact us/professor for an updated code.

2.1 Climate Change and How We Approach The Problem

As the advent of Anthropocene and the unprecedented speed of technological development, the emission of greenhouse gases has been surging, rendering the issue of global warming severe. However, the change of climate is so **subtle** that individuals and industries hardly notice the devastating consequences brought by global warming and still hang on to their destructive behaviors such as factory farming, fossil fuel burning, and driving private cars.

Therefore, it is significant to arouse people's awareness of the alarming trend of global warming.

In this project, we will use the per-day average temperature data of **McGuire Air Force Base** from 1955 to 2010 [1] on National Oceanic and Atmospheric Administration (NOAA) website to justify the warming effect happening in the local region. We will construct the basis matrix used for training, compare the accuracy and authenticity of LASSO and Ridge Regression on multiple basis functions and different parameters, and then justify and criticize the model as need be.

Try Me At this point, we recommend you to pause the reading and visualize the temperature data of our McGuire AFB. You can plot the data in MATLAB file `part1.m` (1.1 and 1.2) or an interactive visualizer here.

2.2 Problem: What feature to select in our model?

Throughout the semester, we have been trying to *find the basis vector* of a set of data and describe their characteristic based on what we found. But when it comes to analyzing the a time series data (such as our temperature data), things get tricky. For example, in a straight forward feature selection problem such as the breast cancer model (in HW8), the effect of a gene is largely proportional to its weight in the resulting weight vector in LASSO model. A researcher will simply pick the top few genes to investigate their effect on Brest cancer, which gives them the direction to employ experiment in a straight forward fasion.

However, as you might have seen in the visualization of the data, it is mainly composed of *temperature in different days*. We are not justifying that some specific day takes account to the trend of warming; rather, we want to investigate the probable *factors* that might contribute to global warming.

So here is the crux of the problem:

What features do we want our model to select and provide us useful information about the trend of local warming?

2.3 Basis Function for Data Fitting

One key catch is that you soon realize the biggest factor of the change of temperature: **seasonal cycle**. At the one-year basis level, the temperature fluctuate regularly with the rhythm of four season, and this cycle is, in fact, the most significant factor that influence the change of temperature.

Mathematically speaking, we can assume a relatively good fit using a sin (or cos) function with a cycle of a year (around 365.25 days, with consideration of leap years).

$$y = w \cdot \sin(\frac{2\pi}{T_{yr}} \cdot (x - \phi))$$

where $T_{yr} = 365.25$.

By convention, it is more favorable to use a pair of sin and cos function to represent this function equivalently as a sum of its orthogonal functions:

$$y = w_1 \cdot \sin(\frac{2\pi}{T_{yr}}x) + w_2 \cdot \cos(\frac{2\pi}{T_{yr}}x)$$

What if we also want to take consideration of the cycle of US Presidential Election (and see if the heat of politics will have an influence of the temperature)?

$$\begin{aligned} y = & w_1 \cdot \sin(\frac{2\pi}{T_1}x) + w_2 \cdot \sin(\frac{2\pi}{T_2}x) + \\ & w_3 \cdot \cos(\frac{2\pi}{T_1}x) + w_4 \cdot \cos(\frac{2\pi}{T_2}x) \end{aligned}$$

Therefore, our function now has 4 orthogonal components (orthogonal functions) represents by different cycles of sin/cos function.

By now you should get a sense of why we employ basis function: they give us the ability to **choose specific features of interest** (the cyclical property of a potential cause of temperature change) and **let the model to learn the weights** of the corresponding function to determine the significance of the feature in the model. By introducing a basis function, we now take a perspective to look at the maximum basis function, and analyze the feature corresponds to that function based on the weight vector we get after the training.

Try Me At this point, we encourage you to pause and look at the visualization of the two models described above in this *interactive example*.

2.4 Constructing A Basis Matrix

Machine learning people love matrix multiplication. It gives them (and hopefully, you) the freedom to directly apply previously established model on the training process. Thus, the next problem is how to convert our data into a simple matrix representation:

$$\mathbf{y} = \mathbf{A}\mathbf{w}$$

For example, this model

$$y = w_1 \cdot \sin(\frac{2\pi}{T_{yr}}x) + w_2 \cdot \cos(\frac{2\pi}{T_{yr}}x)$$

will be expressed as

$$\mathbf{A} = \begin{bmatrix} \sin(2\pi x_1/T_1) & \cos(2\pi x_1/T_1) \\ \sin(2\pi x_2/T_1) & \cos(2\pi x_2/T_1) \\ \vdots & \vdots \\ \sin(2\pi x_n/T_1) & \cos(2\pi x_n/T_1) \end{bmatrix}$$

In a Matlab perspective:

```
x = (1:N)';          % column vector
t = [ (2*pi/T1) ];   % row vector
X = x ./ t;          % outer product of x and 1./t (element-wise inverse)
A = [  sin( X )  cos( X )  ]; % The basis matrix
```

Try Me in fear we sperad the content too lengthy, please try the link here and see a detailed explaination of how the basis matrix is constructed.

3 Warm-up

In this section, we will let you play around with the data and get a sense of feeling about the temperature change of the McGuire AFB in the 55 year span. Here are some questions that might help you through the process.

Problem 1. The file `part1.m` helps you to load the data for practice. Plot the data for 55 years, 11 years, 5 years and 2 years (code already written). What kind of basis function would you want to apply to fit the data?

Problem 2. Try `online material Section 1.2` and `part1.m (Section 1.3)` to play around the model with different parameters. Try fit the 2-year-range data with a combination of different basis functions. Which is better? How do you judge a model is better than others?

- $\sin(x)$
- $\sin(x), \cos(x)$
- $\sin(x), \sin(2x)$

Problem 3. Write down the basis matrix for the aforementioned model:

$$\begin{aligned} y = & w_1 \cdot \sin(\frac{2\pi}{T_1}x) + w_2 \cdot \sin(\frac{2\pi}{T_2}x) + \\ & w_3 \cdot \cos(\frac{2\pi}{T_1}x) + w_4 \cdot \cos(\frac{2\pi}{T_2}x) \end{aligned}$$

Problem 4. Goto sectin 1.5 of `part1.m`. What other factors do you think might contribute to global warming, and what do you think will contribute the most to global warming? Here are some of our thoughts:

- T = 0.50 : Undergrad Final Exam Cycle
- T = 1.00 : Seasonal Cycle
- T = 4.00 : US President Election
- T = 10.78 : Solar Cycle
- T = 18.60 : Moon Declination angle changing cycle

4 Main Activity

The section is largely based on `part2.m`.

Problem 1. Recall the property of LASSO and Ridge Regression. Which model will you choose to apply on the function? Which is better and why?

Problem 2. Run code in Section 2 in `part2.m`. It gives you the sense of how LASSO and Ridge actually performs with the selected set of factors.

1. What is the basis function we applied in the code? Notice that we add a constant offset. Why is it necessary?

2. What does the figure shows and why? Change a set of cycles and run the comparison again. Also run them on different set of regularization parameters. Is there significant differences? Why?

3. You should by now realize that the factors you choose will also have an influence on how the model performs. Can you modify the code to use a larger set of `t` to "find" the good factor? Remember to modify the `t` and be careful of the dimension of the rest of the parts. (If you get an error, it is highly possible you failed in one of the multiplication step).

Problem 3. One extraordinary approach to the verification of global warming comes from the following hypothesis:

"If we extract the global warming induced by (significant) cyclic environment factor, and still get a gradual rise of temperature, then we can say that global warming exist and is indeed induced by activities other than natural causes."

This simplified **proof by contradiction** indicate that we should modify our model to include:

(a). All significant factors that cause by cyclical nature phenomenon.

(b). A linear term that indicate the rising/falling of the temperature.

1. Construct the basis matrix for the model and run LASSO and Ridge similar to the previous problem. What do you observe?

2. Some people say that it is essential to regularize the coefficient of linear term, otherwise the model will suffer from the explosive linear term and get a bad result. What do you think? How do you regularize if it is necessary?

3. In Vanderbei's famous paper *Local Warming or Global Warming*, he analyze the problem based on the optimization problem as follows:

$$\min_{w_0,w_1,\ldots,w_5} \sum_{x=1}^{N} |w_0 + w_1 x + w_2 cos(2\pi x/T_{yr}) + w_3 sin(2\pi x/T_{yr}) + w_4 cos(2\pi x/T_{solar}) + w_5 sin(2\pi x/T_{solar})|$$

where $T_{yr} = 365.25$ and $T_{soloar} = (365.25 \times 10.7)$. This sounds pretty reasonable since the fundamental astrophysics proves that seasonal cycle and solar cycle influence the average daily temperature for most typical part on earth.

In this activity, instead of using the absolute value, we could use the squared error to simplify the problem:

$$\min_{w_0,w_1,\ldots,w_5} \sum_{x=1}^{N} \{w_0 + w_1 x + w_2 cos(2\pi x/T_{yr}) + w_3 sin(2\pi x/T_{yr}) + w_4 cos(2\pi x/T_{solar}) + w_5 sin(2\pi x/T_{solar})\}^2$$

(a) Formulate your code to run on LASSO and Ridge. Observe the weight vector and how each component contribute to the change of temperature.

(b) What does the coefficient of the linear term suggests? Does it predict a reasonable rate of warming (in your knowledge)?

5 Optional Activity

With the constrain of time, this section serves for student who are passionate to go beyond the scope of Local Warming and probably justify the effect of global warming. NOAA provides abundant dataset [3] for users to download, and we have prepare and processed some of them `in our Github repo`.

Here upon we ask something more interesting. We haven't get the full answer of these questions, but you might be the one who find the answer and start your new career.

Problem 1. Verify your model and Robert's model using data from other parts of the world (including Madison). Is the model always predict the warming trend of the globe? What does it suggest?

Problem 2. What does the coefficient of the solar cycle term implies? If it is too insignificant compared to seasonal factor, why would Robert want to include the term in his analysis? If you're curious, refer to his paper and see if you can answer the question. We are very regretful that the content is beyond our scope of discussion...

6 Solution: https://www.overleaf.com/read/vvffjgrfhwcb

The solution of the project in linked **here** : **https://www.overleaf.com/read/vvffjgrfhwcb**

The purpose of an open link solution is for more people to edit and share their thoughts once they finish the activity and seek for a better answer in the open questions. We encourage you to edit the page if you are interested in, and your answer will be of your own credit.

The solution is also provided offline so you could check it in your folder.

7 Reference

1. McGuireAFB Dataset (processed version by Rebert)
2. Vanderbei, R. J. (2012). Local warming. SIAM Review, 54(3), 597-606.
3. NOAA Climate Information Dataset