Assignment #1

CS181-S22

Due: 7:59pm ET, February 4, 2022

Homework 1: Regression

Introduction

This homework is on different forms of linear regression and focuses on loss functions, optimizers, and regularization. Linear regression will be one of the few models that we see that has an analytical solution. These problems focus on deriving these solutions and exploring their properties.

If you find that you are having trouble with the first couple problems, we recommend going over the fundamentals of linear algebra and matrix calculus (see links on website). The relevant parts of the cs181-textbook notes are Sections 2.1 - 2.7. We strongly recommend reading the textbook before beginning the homework.

We also encourage you to first read the Bishop textbook, particularly: Section 2.3 (Properties of Gaussian Distributions), Section 3.1 (Linear Basis Regression), and Section 3.3 (Bayesian Linear Regression). (Note that our notation is slightly different but the underlying mathematics remains the same!).

Please type your solutions after the corresponding problems using this LaTeX template, and start each problem on a new page. You may find the following introductory resources on LaTeX useful: LaTeX Basics and LaTeX tutorial with exercises in Overleaf

Homeworks will be submitted through Gradescope. You will be added to the course Gradescope once you join the course Canvas page. If you haven't received an invitation, contact the course staff through Ed.

Please submit the writeup PDF to the Gradescope assignment 'HW1'. Remember to assign pages for each question.

Please submit your LaTeXfile and code files to the Gradescope assignment 'HW1 - Supplemental'. Your files should be named in the same way as we provide them in the repository, e.g. `T1_P1.py`, etc.

Problem 1 (Optimizing a Kernel, 15pts)

Kernel-based regression techniques are similar to nearest-neighbor regressors: rather than fit a parametric model, they predict values for new data points by interpolating values from existing points in the training set. In this problem, we will consider a kernel-based regressor of the form:

$$f(x^*) = \sum_n K(x_n, x^*) y_n$$

where (x_n, y_n) are the training data points, and $K(x, x')$ is a kernel function that defines the similarity between two inputs x and x'. Assume that each x_i is represented as a column vector, i.e. a D by 1 vector where D is the number of features for each data point. A popular choice of kernel is a function that decays as the distance between the two points increases, such as

$$K(x, x') = \exp\left(\frac{-||x - x'||_2^2}{\tau}\right) = \exp\left(\frac{-(x - x')^T(x - x')}{\tau}\right)$$

where τ represents the square of the lengthscale (a scalar value). In this problem, we will consider optimizing what that (squared) lengthscale should be.

1. Let $\{(x_n, y_n)\}_{n=1}^N$ be our training data set. Suppose we are interested in minimizing the residual sum of squares. Write down this loss over the training data $\mathcal{L}(W)$ as a function of τ.

   Important: When computing the prediction $f(x_i)$ for a point x_i in the training set, carefully consider for which points x' you should be including the term $K(x_i, x')$ in the sum.

2. Take the derivative of the loss function with respect to τ.

1. Since $x_i \in \{x_1, \dots, x_n\}$, we'll define $f(x_i)$ as $\sum_{j=1, j\neq i}^n K(x_j, x_i) y_j$. Thus the loss function is

$$\mathcal{L}(\tau) = \sum_{i=1}^n (f(x_i) - y_i)^2 = \sum_{i=1}^n \left(\left(\sum_{j=1, j\neq i}^n K(x_j, x_i) y_j\right) - y_i\right)^2 = \sum_{i=1}^n \left(\left(\sum_{j=1, j\neq i}^n e^{\frac{-\|x_j - x_i\|_2^2}{\tau}} y_j\right) - y_i\right)^2.$$

2. Taking the derivative with the chain rule yields

$$\frac{d\mathcal{L}}{d\tau} = \sum_{i=1}^n 2\left(\left(\sum_{j=1, j\neq i}^n e^{\frac{-\|x_j - x_i\|_2^2}{\tau}} y_j\right) - y_i\right)\left(\sum_{j=1, j\neq i}^n \left(\frac{\|x_j - x_i\|_2^2}{\tau^2}\right) e^{\frac{-\|x_j - x_i\|_2^2}{\tau}} y_j\right).$$

Problem 1 (cont.)

3. Consider the following data set:

```
x , y
0 , 0
1 , 0.5
2 , 1
3 , 2
4 , 1
6 , 1.5
8 , 0.5
```

   And the following lengthscales: $\tau = .01$, $\tau = 2$, and $\tau = 100$.

   Write some Python code to compute the loss with respect to each kernel for the dataset provided above. Which lengthscale does best? For this problem, you can use our staff **script to compare your code to a set of staff-written test cases.** This requires, however, that you use the structure of the starter code provided in `T1_P1.py`. More specific instructions can be found at the top of the file `T1_P1_Testcases.py`. You may run the test cases in the command-line using `python T1_P1_TestCases.py`. **Note that our set of test cases is not comprehensive: just because you pass does not mean your solution is correct! We strongly encourage you to write your own test cases and read more about ours in the comments of the Python script.**

4. Plot the function $(x^*, f(x^*))$ for each of the lengthscales above. You will plot x^* on the x-axis and the prediction $f(x^*)$ on the y-axis. For the test inputs x^*, you should use an even grid of spacing of 0.1 between $x^* = 0$ and $x^* = 12$. (Note: it is possible that a test input x^* lands right on top of one of the training inputs above. You can still use the formula!)

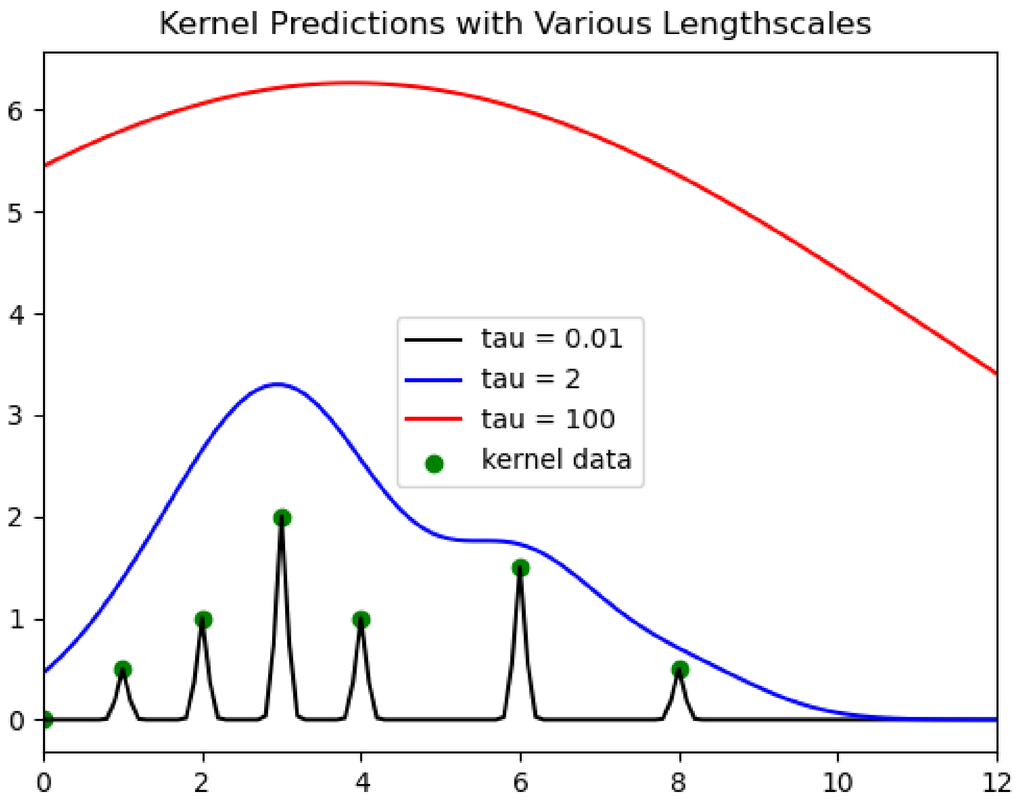
   Initial impressions: Briefly describe what happens in each of the three cases. Is what you see consistent with the which lengthscale appeared to be numerically best above? Describe why or why not.

5. Bonus: Code up a gradient descent to optimize the kernel for the data set above. Start your gradient descent from $\tau = 2$. Report on what you find.

   Note: Gradient descent is discussed in Section 3.4 of the cs181-textbook notes and Section 5.2.4 of Bishop, and will be covered later in the course!

3. The best of the three lengthscales is $\tau = 2$ with a loss of 3.305 while $\tau = 0.01$ has a loss of 8.75 and $\tau = 100$ has a loss of 120.36.

4. Plotting yields:



For $\tau = 0.01$, the graph is mostly flat with values close to 0 but spikes near the x values in our data. At each of the x values from our data, the function is approximately equal to the corresponding y value. If x^* is not very close to one of the x values from our data, then the exponent of each kernel similarity term is a negative number with a large magnitude since we're dividing by $\tau = 0.01$ in the exponent. Thus each term in the summation for $f(x^*)$ ends up being very close to 0, and $f(x^*)$ as a whole is also close to 0.

For $\tau = 2$, the graph roughly follows the shape of the data points but is shifted up with the shift on the left side being larger than the shift on the right.

For $\tau = 100$, the graph resembles a downward-opening, wide parabola that has a vertex around $(4, 6)$. Unless x^* is very far from all of our data points, the exponent in each kernel similarity term will be a negative value with a small magnitude since we're dividing by $\tau = 100$ in the exponent. Thus each kernel similarity term will be closer to 1, and $f(x^*)$ will resemble the sum of all y values from our data for $x^* \in [0, 8]$.

These graphs are consistent with the best lengthscale that we found in part 3 since $\tau = 2$ produces a curve that has a similar shape to our data with a shift that is not too big.

Problem 2 (Kernels and kNN, 10pts)

Now, let us compare the kernel-based approach to an approach based on nearest-neighbors. Recall that kNN uses a predictor of the form

$$f(x^*) = \frac{1}{k} \sum_n y_n \mathbb{I}(x_n \texttt{ is one of k-closest to } x^*)$$

where $\mathbb{I}$ is an indicator variable. For this problem, you will use the **same dataset and kernel as in Problem 1**.

For this problem, you can use our staff **script to compare your code to a set of staff-written test cases.** This requires, however, that you use the structure of the starter code provided in `T1_P2.py`. More specific instructions can be found at the top of the file `T1_P2_Testcases.py`. You may run the test cases in the command-line using `python T1_P2_TestCases.py`. **Note that our set of test cases is not comprehensive: just because you pass does not mean your solution is correct! We strongly encourage you to write your own test cases and read more about ours in the comments of the Python script.**

Make sure to include all required plots in your PDF.

1. Implement kNN for $k = \{1, 3, N-1\}$ where N is the size of the dataset, then plot the results for each k. To find the distance between points, use the kernel function from Problem 1 with lengthscale $\tau = 1$.

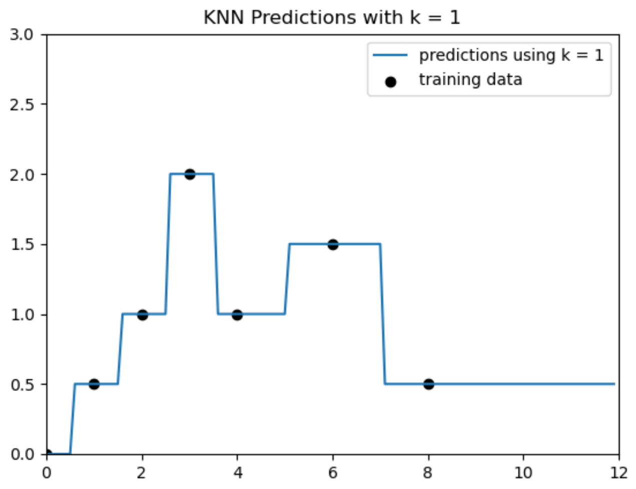
   As before, you will plot x^* on the x-axis and the prediction $f(x^*)$ on the y-axis. For the test inputs x^*, you should use an even grid of spacing of 0.1 between $x^* = 0$ and $x^* = 12$. (Like in Problem 1, if a test point lies on top of a training input, use the formula without excluding that training input.)

   You may choose to use some starter Python code to create your plots provided in `T1_P2.py`. Please **write your own implementation of kNN** for full credit. Do not use external libraries to find nearest neighbors.

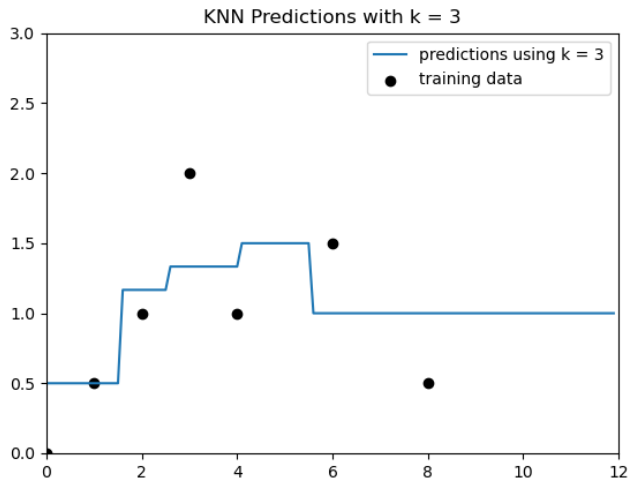
2. Describe what you see: What is the behavior of the functions in these three plots? How does it compare to the behavior of the functions in the three plots from Problem 1? Are there situations in which kNN and kernel-based regression interpolate similarly? Extrapolate similarly? Based on what you see, do you believe there exist some values of k and τ for which the kNN and kernel-based regressors produce the exact same classifier (i.e. given *any* point x, the two regressors will produce the same prediction $f(x)$)? Explain your answer.

3. Why did we not vary τ for the kNN approach?

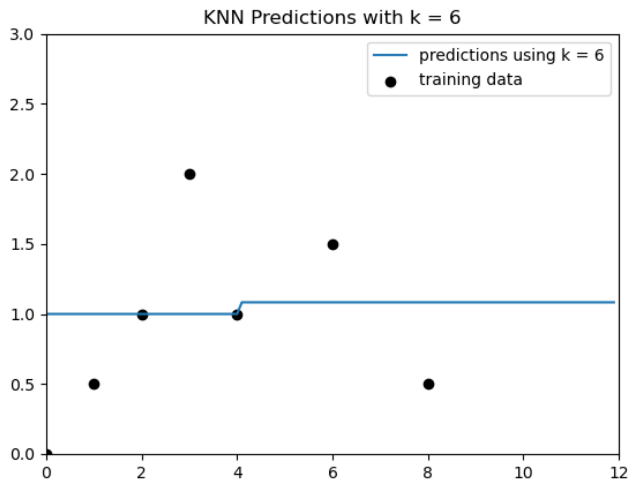
1. The kNN regressions are:


KNN Predictions with k = 1
predictions using k = 1
training data
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12

i.


KNN Predictions with k = 3
predictions using k = 3
training data
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12

ii.



iii.

2. When $k = 1$, the regression only considers the closest x_i value, so the value of $f(x^*)$ is dominated entirely by this value. This is similar to the kernel based approach with $\tau = 0.01$; however, in this case, $f(x^*)$ will still resemble a y_i instead of being almost 0. When $k = N - 1$, the regression takes the average of all but one of the y_i values. Thus the curve ends up being very flat with only two values for $f(x^*)$. This is similar to the kernel based approach with $\tau = 100$ where all y_i values were weighted heavily, but it is far more rigid. When $k = 3$, the regression considers the closest 3 values. While this is the Goldilocks method for this problem, the curve does not seem as accurate as the curve from the kernel based approach with $\tau = 2$, and it's not smooth either.

   The two approaches do not interpolate similarly with the closest pair being $k = 3$ for kNN and $\tau = 2$. They do not extrapolate similarly either since all of the kernel based regressions quickly converge toward 0 on either side of our data while the kNN regressions converge to 0.5 or around 1 on either side of our data.

   I do not think the two approaches can ever produce the exact same classifier. We can think of $f(x^*)$ as N functions that correspond to the coefficient for each y_n in the summation (e.g. for the kernel, this would be $K(x^*, x_n)$). In order for the classifiers to be the same, these functions would have to be the same, and the kernel functions are smooth while the kNN functions are not. Thus they cannot ever be the same classifier.

3. There is no need to vary τ since it simply scales the exponent. Let $z_n = \|x_n - x^*\|_2^2$ for all $n \in [1, N]$. If x_n is one of the k closest values to x^*, then z_n will be one of the k smallest values in $\{z_1, \ldots, z_N\}$. It follows that $e^{-\frac{z_n}{\tau}}$ will be one of the k greatest values in $\{e^{-\frac{z_1}{\tau}}, \ldots, e^{-\frac{z_N}{\tau}}\}$ (as long as τ is positive) since $e^{-\frac{z_n}{\tau}}$ increases as z_n decreases and vice versa. Thus there is no need to vary τ since the kernel function ranking will always return the same k values (for a particular x^*) as long as τ is positive. Intuitively, the k closest values to x^* do not depend on τ, and the kernel function just gives us a way to determine the k closest values.

Problem 3 (Deriving Linear Regression, 10pts)

The solution for the least squares linear regressions "looks" kind of like a ratio of covariance and variance terms. In this problem, we will make that connection more explicit.

Let us assume that our data are tuples of scalars (x, y) that are described by some joint distribution $p(x, y)$. For clarification, the joint distribution $p(x, y)$ is just another way of saying the "joint PDF" $f(x, y)$, which may be more familiar to those who have taken Stat 110, or equivalent.

We will consider the process of fitting these data from this distribution with the best linear model possible, that is a linear model of the form $\hat{y} = wx$ that minimizes the expected squared loss $E_{x,y}[(y - \hat{y})^2]$.

Notes: The notation $E_{x,y}$ indicates an expectation taken over the joint distribution $p(x, y)$. Since x and y are scalars, w is also a scalar.

1. Derive an expression for the optimal w, that is, the w that minimizes the expected squared loss above. You should leave your answer in terms of moments of the distribution, e.g. terms like $E_x[x]$, $E_x[x^2]$, $E_y[y]$, $E_y[y^2]$, $E_{x,y}[xy]$ etc.
2. Provide unbiased and consistent formulas to estimate $E_{x,y}[yx]$ and $E_x[x^2]$ given observed data $\{(x_n, y_n)\}_{n=1}^N$.
3. In general, moment terms like $E_{x,y}[yx]$, $E_{x,y}[x^2]$, $E_{x,y}[yx^3]$, $E_{x,y}[\frac{x}{y}]$, etc. can easily be estimated from the data (like you did above). If you substitute in these empirical moments, how does your expression for the optimal w^* in this problem compare with the optimal w^* that we see in Section 2.6 of the cs181-textbook?
4. Many common probabilistic linear regression models assume that variables x and y are jointly Gaussian. Did any of your above derivations rely on the assumption that x and y are jointly Gaussian? Why or why not?

1. We expand $E_{x,y}[(y - \hat{y})^2]$ as follows:

$$\begin{aligned} E_{x,y}[(y-\hat{y})^2] &= E_{x,y}[y^2 - 2y\hat{y} + \hat{y}^2] \\ &= E_{x,y}[y^2] - E_{x,y}[2y\hat{y}] + E_{x,y}[\hat{y}^2] \\ &= E_y[y^2] - 2E_{x,y}[ywx] + E_{x,y}[(wx)^2] \\ &= E_y[y^2] - 2wE_{x,y}[xy] + w^2 E_x[x^2] \end{aligned}$$

   Since this expression is a quadratic in terms of w and the leading coefficient $E_x[x^2]$ is positive (as long as x is not always 0), we can minimize the expected loss by taking the vertex of the quadratic. Therefore

$$w^* = \frac{-(-2E_{x,y}[xy])}{2E_x[x^2]} = \frac{E_{x,y}[xy]}{E_x[x^2]}.$$

2. It was stated on Ed (thread #9) that we can assume that the data is drawn independently from the joint distribution, so by the law of large numbers, we can approximate $E_{x,y}[xy]$ with $\frac{1}{N}\sum_{n=1}^N x_n y_n$ and $E_x[x^2]$ with $\frac{1}{N}\sum_{n=1}^N x_n^2$.

3. Substituting with the empirical moments yields $w^* = \frac{\frac{1}{N}\sum_{n=1}^N x_n y_n}{\frac{1}{N}\sum_{n=1}^N x_n^2} = \frac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n^2}$. This is equivalent to the optimal w^* from the textbook: $(X^T X)^{-1} X^T y$. Since our inputs x_n are scalars, $X^T X = \sum_{n=1}^N x_n^2$

and $X^T y = \sum_{n=1}^N x_n y_n$. Thus

$$(X^T X)^{-1} X^T y = \frac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n^2}.$$

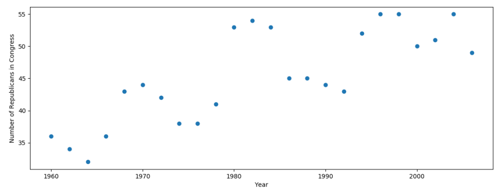
4. I don't think so since all of the manipulations used general properties of expected value and independent draws from identical distributions rather than properties that only apply to the multivariate normal.

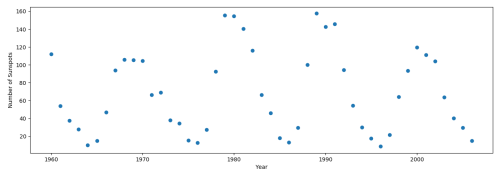
Problem 4 (Modeling Changes in Republicans and Sunspots, 15pts)

The objective of this problem is to learn about linear regression with basis functions by modeling the number of Republicans in the Senate. The file `data/year-sunspots-republicans.csv` contains the data you will use for this problem. It has three columns. The first one is an integer that indicates the year. The second is the number of Sunspots observed in that year. The third is the number of Republicans in the Senate for that year. The data file looks like this:

```
Year,Sunspot_Count,Republican_Count
1960,112.3,36
1962,37.6,34
1964,10.2,32
1966,47.0,36
```

You can see scatterplots of the data in the figures below. The horizontal axis is the Year, and the vertical axis is the Number of Republicans and the Number of Sunspots, respectively.





(Data Source: http://www.realclimate.org/data/senators_sunspots.txt)

Make sure to include all required plots in your PDF.

1. In this problem you will implement ordinary least squares regression using 4 different basis functions for **Year (x-axis)** v. **Number of Republicans in the Senate (y-axis)**. Some starter Python code that implements simple linear regression is provided in `T1_P4.py`.

   Note: The numbers in the *Year* column are large (between 1960 and 2006), especially when raised to various powers. To avoid numerical instability due to ill-conditioned matrices in most numerical computing systems, we will scale the data first: specifically, we will scale all "year" inputs by subtracting 1960 and then dividing by 40. Similarly, to avoid numerical instability with numbers in the *Sunspot_Count* column, we will also scale the data first by dividing all "sunspot count" inputs by 20. Both of these scaling procedures have already been implemented in lines $65-69$ of the starter code in `T1_P4.py`. Please do *not* change these lines!

   First, plot the data and regression lines for each of the following sets of basis functions, and include the generated plot as an image in your submission PDF. You will therefore make 4 total plots:

   (a) $\phi_j(x) = x^j$ for $j = 1, \ldots, 5$
   ie, use basis $y = a_1x^1 + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$ for some constants $\{a_1, ..., a_5\}$.

   (b) $\phi_j(x) = \exp \frac{-(x-\mu_j)^2}{25}$ for $\mu_j = 1960, 1965, 1970, 1975, \ldots 2010$

   (c) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 5$

   (d) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 25$

   * Note: Please make sure to add a bias term for all your basis functions above in your implementation of the `make_basis` function in `T1_P4.py`.

   Second, for each plot include the residual sum of squares error. Submit the generated plot and residual sum-of-squares error for each basis in your LaTeX write-up.

Problem 4 (cont.)

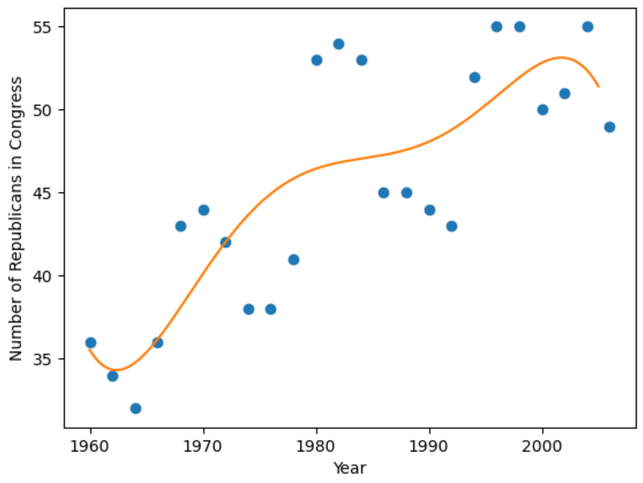
2. Repeat the same exact process as above but for **Number of Sunspots (x-axis)** v. **Number of Republicans in the Senate (y-axis)**. Now, however, only use data from before 1985, and only use basis functions (a), (c), and (d) – ignore basis (b). You will therefore make 3 total plots. For each plot make sure to also include the residual sum of squares error.

   Which of the three bases (a, c, d) provided the "best" fit? **Choose one**, and keep in mind the generalizability of the model.

   Given the quality of this fit, do you believe that the number of sunspots controls the number of Republicans in the senate (Yes or No)?

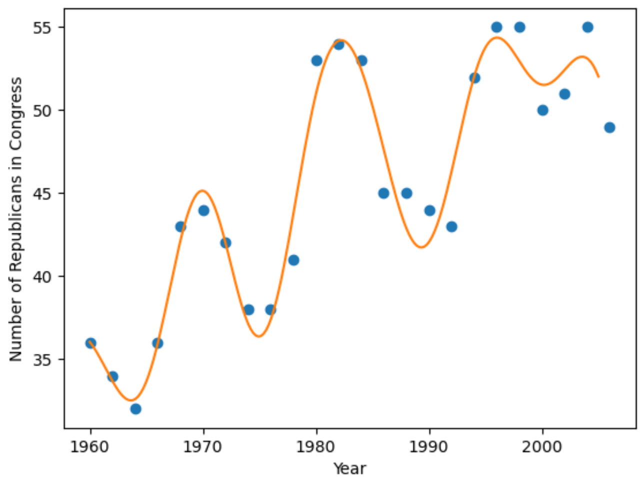
1. Using the given bases with the year as input yields the following regressions:

   (a) The part a basis yields the following regression:



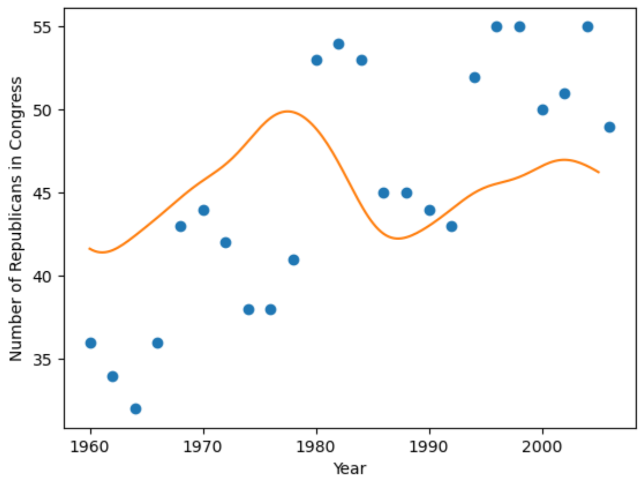
   The residual sum of squares error for the part a basis is approximately 395.

   (b) The part b basis yields the following regression:



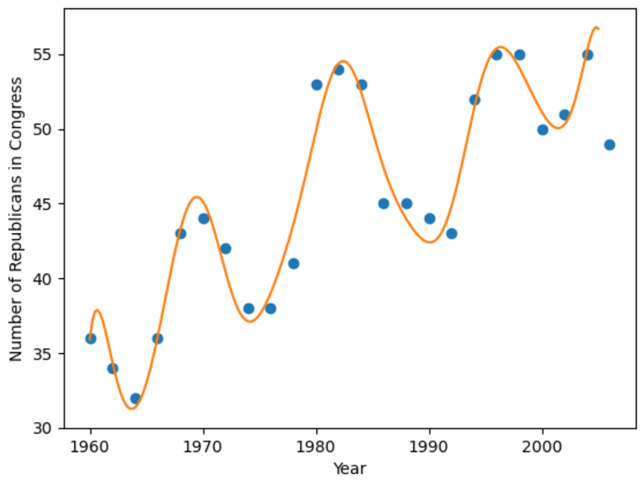
The residual sum of squares error for the part b basis is approximately 54.3.

(c) The part c basis yields the following regression:



The residual sum of squares error for the part c basis is approximately 1082.8.

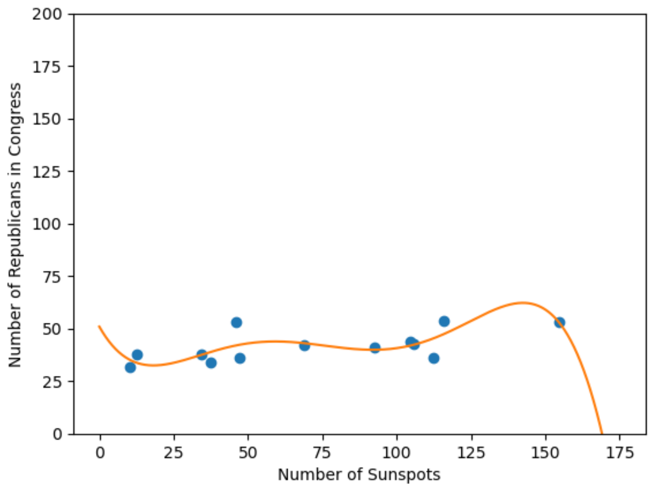
(d) The part d basis yields the following regression:



The residual sum of squares error for the part d basis is approximately 39.

2. Using the given bases with the number of sunspots as input yields the following regressions:

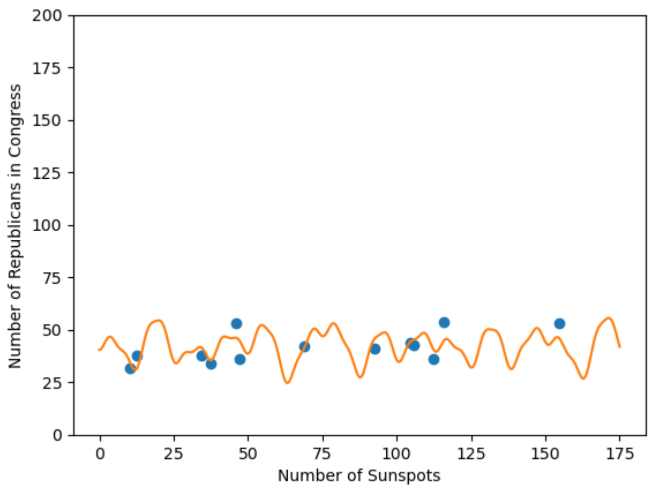
   (a) The part a basis yields the following regression:



The residual sum of squares error for the part a basis is approximately 351.

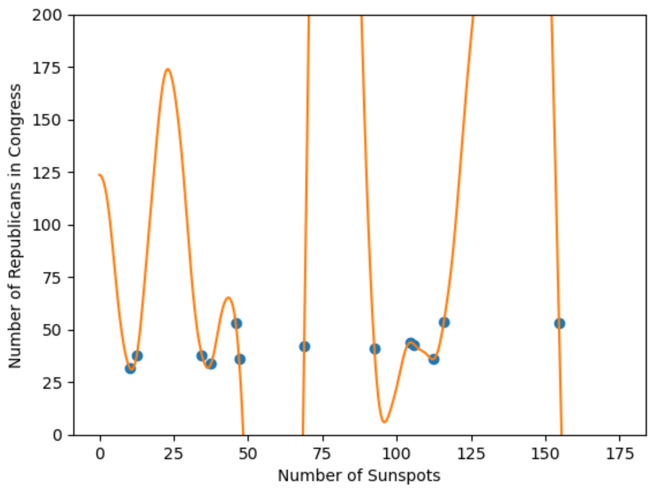
   (b) N/A

   (c) The part c basis yields the following regression:



The residual sum of squares error for the part c basis is approximately 375.1.

(d) The part d basis yields the following regression:



The residual sum of squares error for the part d basis is approximately 0.

I think that the part c provided the best fit. While part a had a smaller loss, it has a sudden drop off at around $x = 160$. While the given data does not contain any data points with x values greater than 160, the shape of the rest of the data does not suggest that such a huge drop off would occur. Part d had a significantly lower loss, but the regression ranges wildly between the data points provided and does not seem to accurately interpolate. While part c does also vary, it varies in a much tighter bound and resembles the overall shape of the data more closely. Thus I think it is the most generalizable model. Based on the fit, I do not think the number of sunspots controls the number of Republicans since the regression just isn't that close, and the data does not appear to be correlated.

Name

Christy Jestin

Collaborators and Resources

I worked with David Qian, and I used Numpy documentation as a reference.

Calibration

This pset took about 10-12 hours.