Assignment #1

Due: 11:59pm February 8, 2019

Homework 1: Linear 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. 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 will submit your solution PDF, your tex file, and your code to Canvas.

Problem 1 (Priors as Regularization,15pts)

In this problem we consider a model of Bayesian linear regression. Define the prior on the parameters as,

$$p(\theta) = \mathcal{N}(\theta \,|\, \mathbf{0}, \sigma_\theta^2 \mathbf{I}),$$

where σ_θ^2 is as scalar variance hyperparameter that controls the variance of the Gaussian prior. Define the likelihood as,

$$p(\mathbf{y} \,|\, \mathbf{X}, \theta) = \prod_{i=1}^{n} \mathcal{N}(y_i \,|\, \theta^\mathsf{T} \mathbf{x}_i, \sigma_n^2),$$

where σ_n^2 is another fixed scalar defining the variance.

1. Using the fact that the posterior is the product of the prior and the likelihood (up to a normalization constant), i.e.,

   $$\arg\max_\theta \ln p(\theta \,|\, \mathbf{y}, \mathbf{X}) = \arg\max_\theta \ln p(\theta) + \ln p(\mathbf{y} \,|\, \mathbf{X}, \theta).$$

   Show that maximizing the log posterior is equivalent to minimizing a regularized loss function given by $\mathcal{L}(\theta) + \lambda \mathcal{R}(\theta)$, where

   $$\mathcal{L}(\theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta^\mathsf{T} \mathbf{x}_i)^2$$
   $$\mathcal{R}(\theta) = \frac{1}{2} \theta^\mathsf{T} \theta$$

   Do this by writing $\ln p(\theta \,|\, \mathbf{y}, \mathbf{X})$ as a function of $\mathcal{L}(\theta)$ and $\mathcal{R}(\theta)$, dropping constant terms if necessary. Conclude that maximizing this posterior is equivalent to minimizing the regularized error term given by $\mathcal{L}(\theta) + \lambda \mathcal{R}(\theta)$ for a λ expressed in terms of the problem's constants.

2. Notice that the form of the posterior is the same as the form of the ridge regression loss

   $$\mathcal{L}(\theta) = (\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta) + \lambda \theta^\top \theta.$$

   Compute the gradient of the loss above with respect to θ. Simplify as much as you can for full credit. Make sure to give your answer in vector form.

3. Suppose that $\lambda > 0$. Knowing that $\mathcal{L}$ is a convex function of its arguments, conclude that a global optimizer of $\mathcal{L}(\theta)$ is

   $$\theta = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \tag{1}$$

   For this part of the problem, assume that the data has been centered, that is, pre-processed such that $\frac{1}{n} \sum_{i=1}^{n} x_{ij} = 0$.

4. What might happen if the number of weights in θ is greater than the number of data points N? How does the regularization help ensure that the inverse in the solution above can be computed?

Solution

1. Since $p(\theta \mid \mathbf{y}, \mathbf{X}) \propto p(\theta)p(\mathbf{y} \mid \mathbf{X}, \theta)$, we can maximize the posterior by maximizing the expression on the right. We have that

$$\begin{aligned} p(\theta)p(\mathbf{y} \mid \mathbf{X}, \theta) &= \mathcal{N}(\theta \mid \mathbf{0}, \sigma_\theta^2 \mathbf{I}) \cdot \prod_{i=1}^{n} \mathcal{N}(y_i \,|\, \theta^\top \mathbf{x}_i, \sigma_n^2) \\ &= \frac{1}{\sqrt{\det 2\pi\sigma_\theta^2 \mathbf{I}}} \exp\{-\tfrac{1}{2}\theta^T(\sigma_\theta^2 \mathbf{I}\theta)^{-1}\} \cdot \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\{-\frac{(y_i - \theta^T x_i)^2}{2\sigma_n^2}\} \end{aligned}$$

Now we take the log (increasing monotonic function, so preserves extrema) and drop additive constants:

$$\ln p(\theta)p(\mathbf{y} \mid \mathbf{X}, \theta) = \ln \frac{1}{\sqrt{\det 2\pi(\sigma_\theta^2 \mathbf{I})^{-1}}} - \tfrac{1}{2}\theta^T(\sigma_\theta^2 \mathbf{I})^{-1}\theta + \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi\sigma_n^2}} - \frac{(y_i - \theta^T x_i)^2}{2\sigma_n^2} \right)$$

$$-\tfrac{1}{2}\theta^T(\sigma_\theta^2 \mathbf{I})^{-1}\theta - \sum_{i=1}^{n} \frac{(y_i - \theta^T x_i)^2}{2\sigma_n^2} = -\frac{1}{2\sigma_\theta^2}\theta^T\theta - \frac{1}{2\sigma_n^2}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$

Maximizing this expression is equivalent to minimizing the negative of this expression scaled up by a factor of σ_n^2, so we aim to minimize

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2 + \frac{\sigma_n^2}{\sigma_\theta^2}\frac{1}{2}\theta^T\theta = \mathcal{L}(\theta) + \lambda\mathcal{R}(\theta)$$

where $\lambda = \frac{\sigma_n^2}{\sigma_\theta^2}$.

2. We use the properties of vector calculus to take the gradient:

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta((\mathbf{y} - \mathbf{X}\theta)^\top(\mathbf{y} - \mathbf{X}\theta) + \lambda\theta^\top\theta)) \\ &= 2(-\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\theta) + 2\lambda\theta \\ &= \boxed{2\mathbf{X}^T\mathbf{X}\theta - 2\mathbf{X}^T\mathbf{y} + 2\lambda\theta} \end{aligned}$$

3. The convexity of $\mathcal{L}$ implies that any optimizer (point where gradient = 0) is necessarily a global optimum. Thus we can set the gradient to 0 and solve for θ:

$$\nabla_\theta L(\theta) = 2\mathbf{X}^T\mathbf{X}\theta - 2\mathbf{X}^T\mathbf{y} + 2\lambda\theta = 0$$

$$\mathbf{X}^T\mathbf{X}\theta + \lambda\theta = \mathbf{X}^T\mathbf{y}$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\theta = \mathbf{X}^T\mathbf{y}$$

$$\theta = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

4. If there are more weights than data points, the matrix $X^T X$ will not have full rank and therefore be

singular/non-invertible. As a result, there could be a set of weights that causes the model to perfectly fit the data, which results in severe overfitting. By adjusting the regularizer λ accordingly, we can make sure the inverse above can be taken.

Problem 2 (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:

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

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'. A popular choice of kernel is a function that decays with the distance between the two points, such as

$$K(x, x') = \exp(-||x - x'||_2^2) = \exp(-(x - x')(x - x')^T)$$

However, the squared Euclidean distance $||x - x'||_2^2$ may not always be the right choice. In this problem, we will consider optimizing over squared Mahalanobis distances

$$K(x, x') = \exp(-(x - x')W(x - x')^T)$$

where W is a symmetric D by D matrix. Intuitively, introducing the weight matrix W allows for different dimensions to matter differently when defining similarity.

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

2. In the following, let us assume that $D = 2$. That means that W has three parameters: W_{11}, W_{22}, and $W_{12} = W_{21}$. Expand the formula for the loss function to be a function of these three parameters.

3. Derive the gradients with respect to each of the parameters in W.

4. Consider the following data set:

```
x1 , x2 , y
  0 , 0 , 0
  0 , .5 , 0
  0 , 1 , 0
  .5 , 0 , .5
  .5 , .5 , .5
  .5 , 1 , .5
  1 , 0 , 1
  1 , .5 , 1
  1 , 1 , 1
```

   And the following kernels:

$$W_1 = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad W_2 = \alpha \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix} \qquad W_3 = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

   with $\alpha = 10$. Write some code to compute the loss with respect to each kernel. Which does best? Why? Does the choice of α matter?

5. **Bonus, ungraded.** Code up a gradient descent to optimize the kernel for the data set above. Start your gradient descent from W_1. Report on what you find.

Solution

1. Our loss function is given by

$$\mathcal{L}(W) = \sum_{i=1}^{N} (y_i - f(x^*))^2,$$

where

$$f(x^*) = \frac{\sum_{n=1}^{N} K(x_n, x^*) y_n}{\sum_{n=1}^{N} K(x_n, x^*)} = \frac{\sum_{n=1}^{N} \exp(-(x_n - x^*)W(x_n - x^*)^T) y_n}{\sum_{n=1}^{N} \exp(-(x_n - x^*)W(x_n - x^*)^T)}$$

so

$$\mathcal{L}(W) = \sum_{i=1}^{N} \left(y_i - \frac{\sum_{n=1}^{N} \exp(-(x_n - x^*)W(x_n - x^*)^T) y_n}{\sum_{n=1}^{N} \exp(-(x_n - x^*)W(x_n - x^*)^T)} \right)^2$$

2. We have that

$$\begin{aligned}\sum_{n=1}^{N} \exp\left(-(x_n - x^*)W(x_n - x^*)^T\right) &= \sum_{n=1}^{N} \exp\left(-[x_1 - x_1^* \; x_2 - x_2^*] \begin{bmatrix} W_{11} & W_{12} \\ W_{12} & W_{22} \end{bmatrix} \begin{bmatrix} x_1 - x_1^* \\ x_2 - x_2^* \end{bmatrix} \right) \\ &= \sum_{n=1}^{N} \exp\left(- \left(w_{11}(x_1 - x_1^*)^2 + w_{22}(x_2 - x_2^*)^2 + 2w_{12}(x_1 - x_1^*)(x_2 - x_2^*) \right) \right)\end{aligned}$$

So

$$\mathcal{L}(W) = \sum_{i=1}^{N} \left(y_i - \frac{\sum_{n=1}^{N} \exp\left(- \left(w_{11}(x_1 - x_1^*)^2 + w_{22}(x_2 - x_2^*)^2 + 2w_{12}(x_1 - x_1^*)(x_2 - x_2^*) \right) \right) y_n}{\sum_{n=1}^{N} \exp\left(- \left(w_{11}(x_1 - x_1^*)^2 + w_{22}(x_2 - x_2^*)^2 + 2w_{12}(x_1 - x_1^*)(x_2 - x_2^*) \right) \right)} \right)^2$$

3. Write $B = (x_1 - x_1^*)^2$, $C = w_{22}(x_2 - x_2^*)^2 + 2w_{12}(x_1 - x_1^*)(x_2 - x_2^*)$. Deriving with respect to w_{11}, we have

$$\begin{aligned}\nabla_{w_{11}} \mathcal{L}(W) &= \sum_{i=1}^{N} \nabla_{w_{11}} \left(y_i - \frac{\sum_{n=1}^{N} \exp(-w_{11}B - C) y_n}{\sum_{n=1}^{N} \exp(-w_{11}B - C)} \right)^2 \\ &= \sum_{i=1}^{N} 2 \left(y_i - \frac{\sum_{n=1}^{N} \exp(-w_{11}B - C) y_n}{\sum_{n=1}^{N} \exp(-w_{11}B - C)} \right) \nabla_{w_{11}} \left(y_i - \frac{\sum_{n=1}^{N} \exp(-w_{11}B - C) y_n}{\sum_{n=1}^{N} \exp(-w_{11}B - C)} \right) \\ &= \sum_{i=1}^{N} 2 \left(y_i - \frac{\sum_{n=1}^{N} \exp(-w_{11}B - C) y_n}{\sum_{n=1}^{N} \exp(-w_{11}B - C)} \right) \cdot\end{aligned}$$

$$\left(\frac{[\sum_{n=1}^{N} \exp(-w_{11}B - C)(By_n)][\sum_{n=1}^{N} \exp(-w_{11}B - C)] - [\sum_{n=1}^{N} \exp(-w_{11}B - C) y_n][\sum_{n=1}^{N} \exp(-w_{11}B - C)(B)]}{(\sum_{n=1}^{N} \exp(-w_{11}B - C))^2} \right)$$

Now write $D = (x_2 - x_2^*)^2$ and $E = w_{11}(x_1 - x_1^*)^2 + 2w_{12}(x_1 - x_1^*)(x_2 - x_2^*)$. In a manner analogous to finding the gradient for $w_1 1$, we get that

$$\nabla_{w_{22}} \mathcal{L}(W) = \sum_{i=1}^{N} 2 \left(y_i - \frac{\sum_{n=1}^{N} \exp(-w_{22}D - E) y_n}{\sum_{n=1}^{N} \exp(-w_{22}D - E)} \right) \cdot$$

$$\left(\frac{[\sum_{n=1}^{N} \exp(-w_{22}D - E)(Dy_n)][\sum_{n=1}^{N} \exp(-w_{22}D - E)] - [\sum_{n=1}^{N} \exp(-w_{22}D - E) y_n][\sum_{n=1}^{N} \exp(-w_{22}D - E)(D)]}{(\sum_{n=1}^{N} \exp(-w_{22}D - E))^2} \right)$$

Finally, write $F = 2(x_1 - x_1^*)(x_2 - x_2^*)$ and $G = w_{11}(x_1 - x_1^*)^2 + w_{22}(x_2 - x_2^*)^2$. In an analogous manner, we get that

$$\nabla_{w_{12}} \mathcal{L}(W) = \sum_{i=1}^{N} 2 \left(y_i - \frac{\sum_{n=1}^{N} \exp(-w_{12}F - G) y_n}{\sum_{n=1}^{N} \exp(-w_{12}F - G)} \right) \cdot$$

$$\left(\frac{[\sum_{n=1}^{N} \exp(-w_{12}F - G)(F y_n)][\sum_{n=1}^{N} \exp(-w_{12}F - G)] - [\sum_{n=1}^{N} \exp(-w_{12}F - G) y_n][\sum_{n=1}^{N} \exp(-w_{22}F - G)(F)]}{(\sum_{n=1}^{N} \exp(-w_{12}F - G))^2} \right)$$

4. We find that $L(W_1) = .169$, $L(W_2) = 1.113$, and $L(W_3) = .012$. It makes sense that W_3 performs best because W_3 weights heavily in favor of x_1, which in the training data seems perfectly correlated with y, as they are always equal. In comparison, x_2 and y seem to have negligible correlation. While the choice of α changes the value of the loss function, it does not seem to be very important in judging the quality of a weight matrix compared to the relative weights assigned to each feature.

Problem 3 (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. The third is the number of Republicans in the Senate. The data file looks like this:

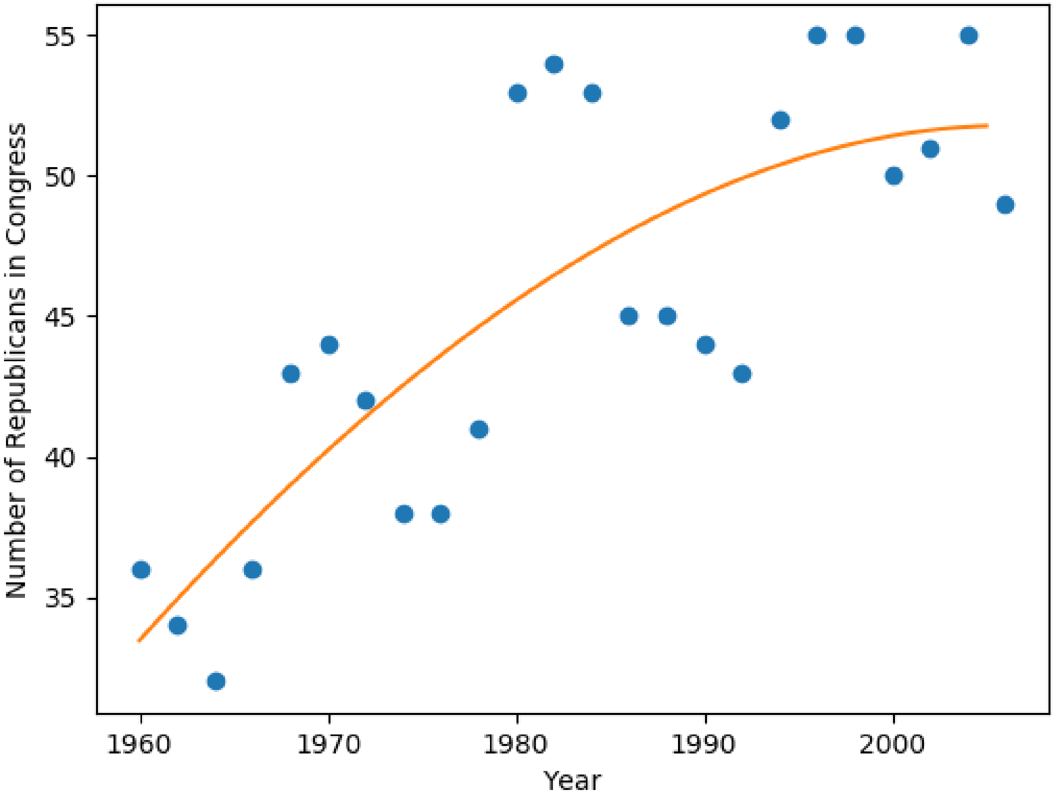
```
Year,Sunspot_Count,Republican_Count
1960,112.3,36
1962,37.6,34
1964,10.2,32
1966,47.0,36
```

and you can see plots 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.

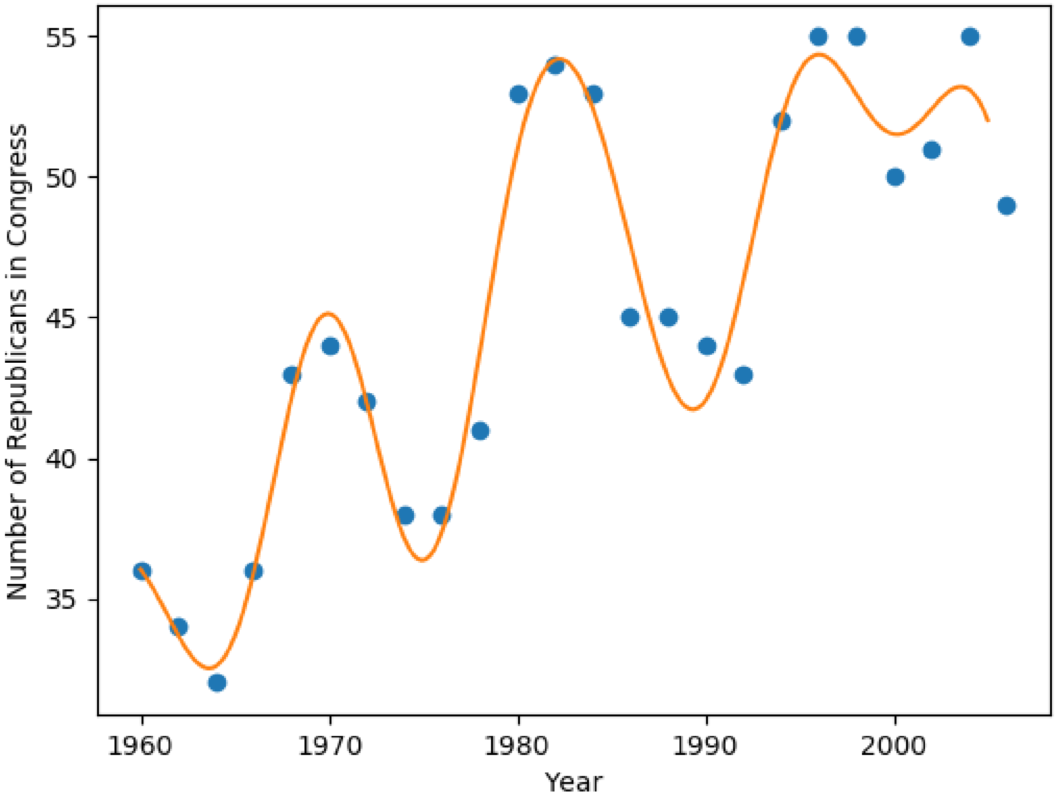
(Data Source: http://www.realclimate.org/data/senators_sunspots.txt)

1. Implement basis function regression with ordinary least squares for years vs. number of Republicans in the Senate. 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 (use basis (b) only for Republicans v. Years, skip for Sunspots v. Years):
   (a) $\phi_j(x) = x^j$ for $j = 1, \ldots, 5$
   (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 $j = 1, \ldots, 25$
2. In addition to the plots, provide one or two sentences for each with numerical support, explaining whether you think it is fitting well, overfitting or underfitting. If it does not fit well, provide a sentence explaining why. A good fit should capture the most important trends in the data.
3. Next, do the same for the number of sunspots vs. number of Republicans, using data only from before 1985. What bases provide the best fit? Given the quality of the fit, would you believe that the number of sunspots controls the number of Republicans in the senate?

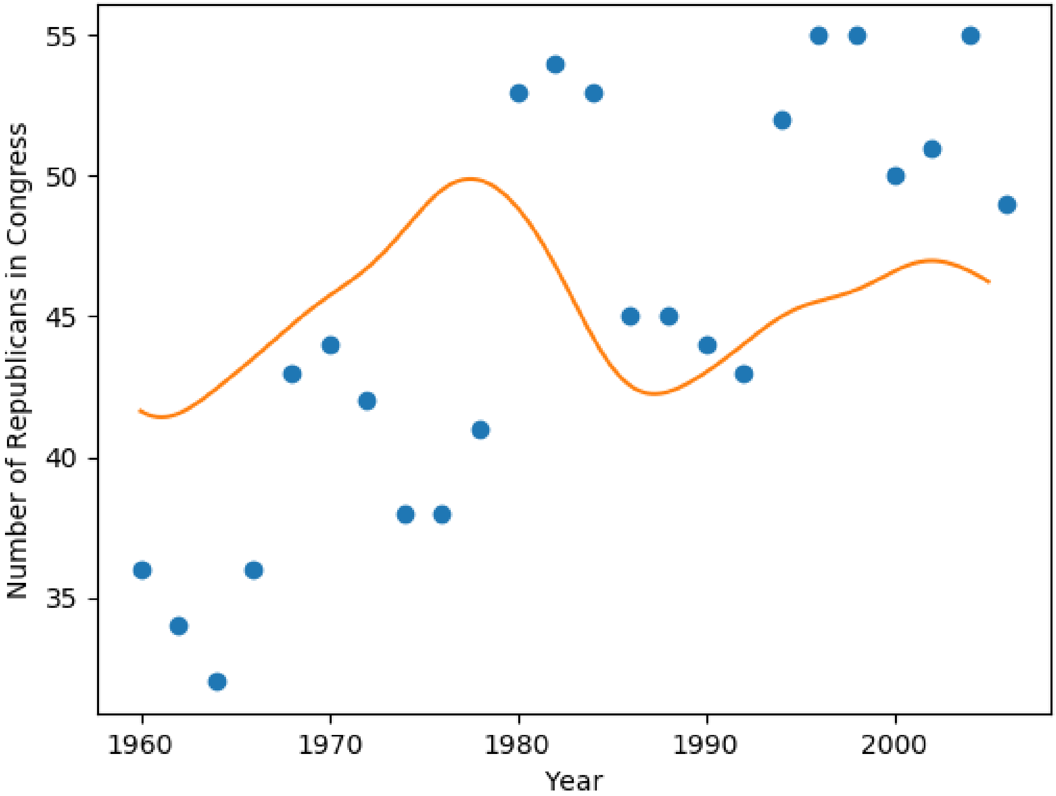
Solution



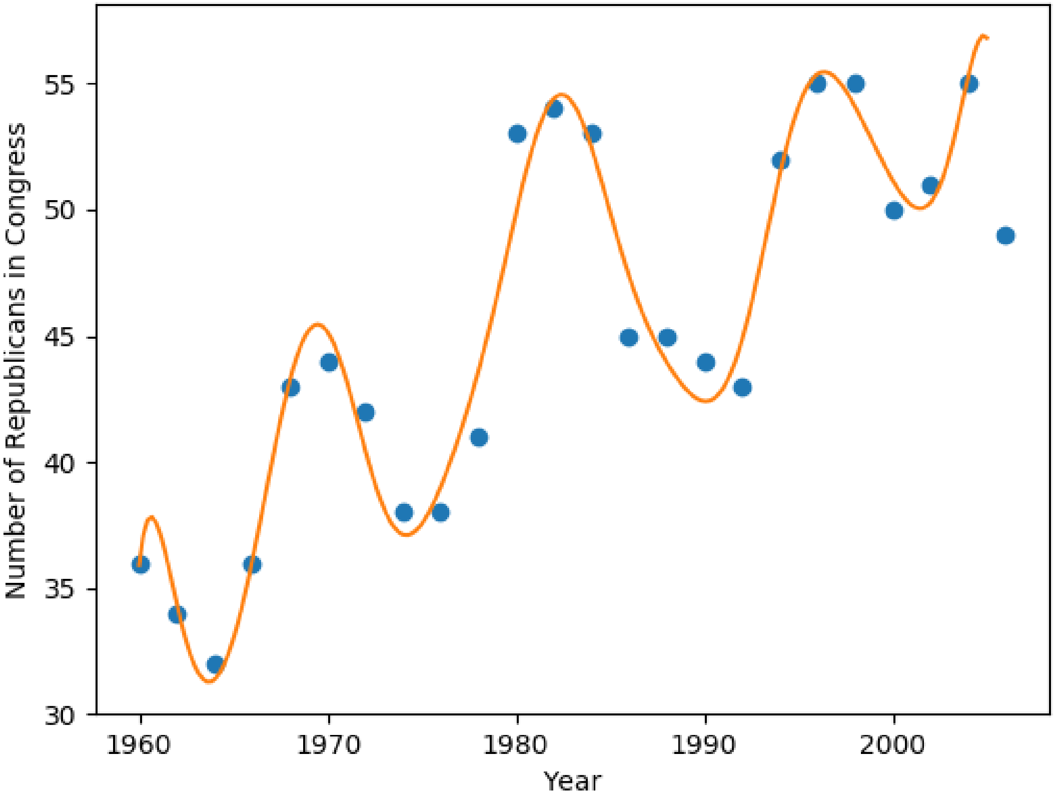
(a) This regression has a loss function value of 212.435. While the model captures the general upward trend of the number of Republicans in Congress, it does not quite capture the oscillatory nature of the data, so it is noticeably underfitting.



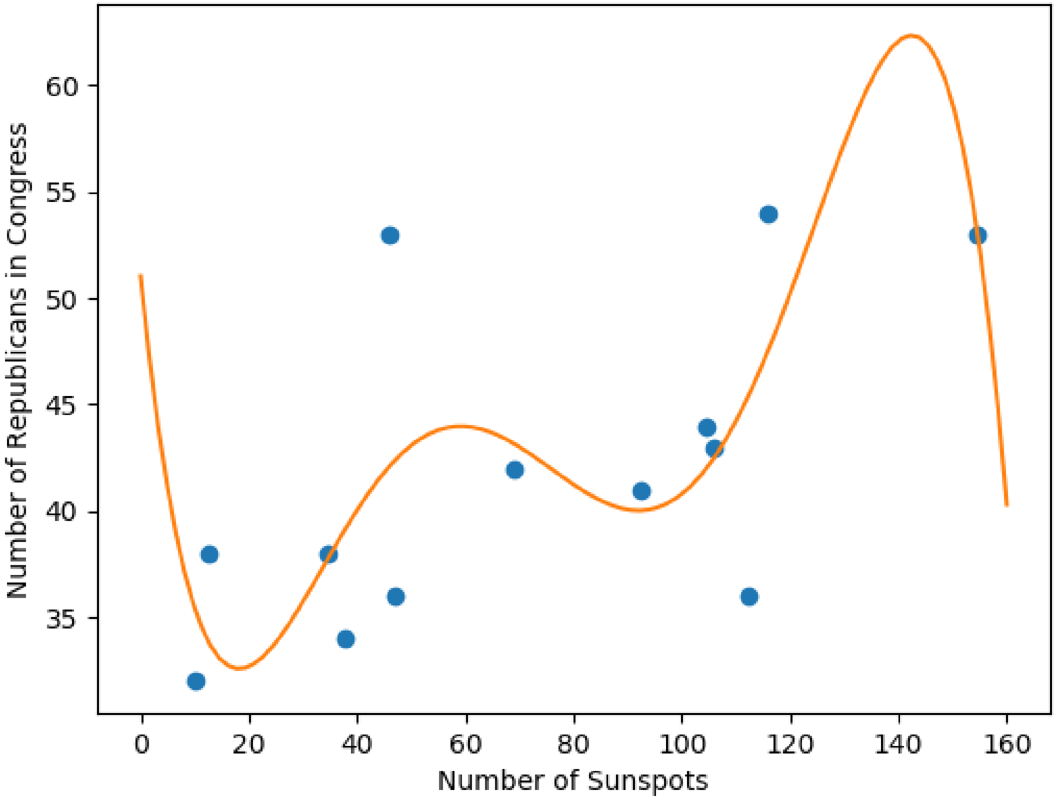
(b) This regression has a loss function value of 27.136. The model captures the general upward trend as well as the oscillatory pattern of the data well, perhaps overfitting slightly but not too much.



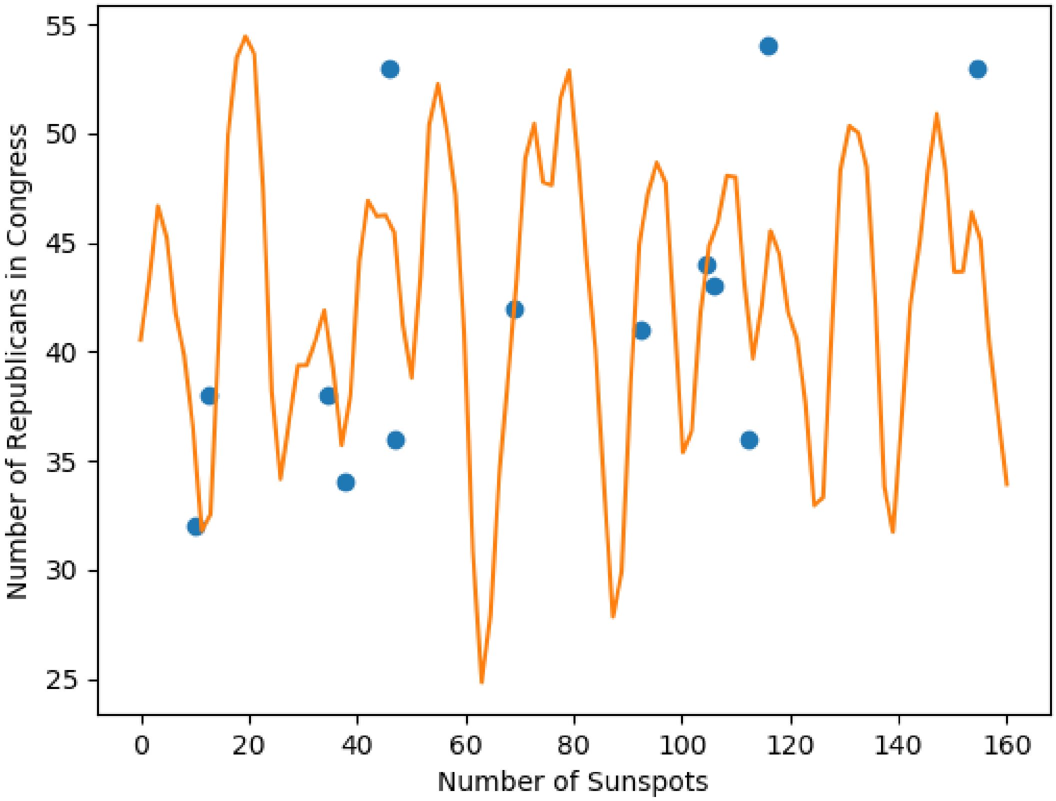
(c) This regression has a loss function value of 541.404. With an inflexible basis function, this model does not reflect the upward trend whatsoever and barely encapsulates the oscillating pattern in the data, so it is severely underfitting.



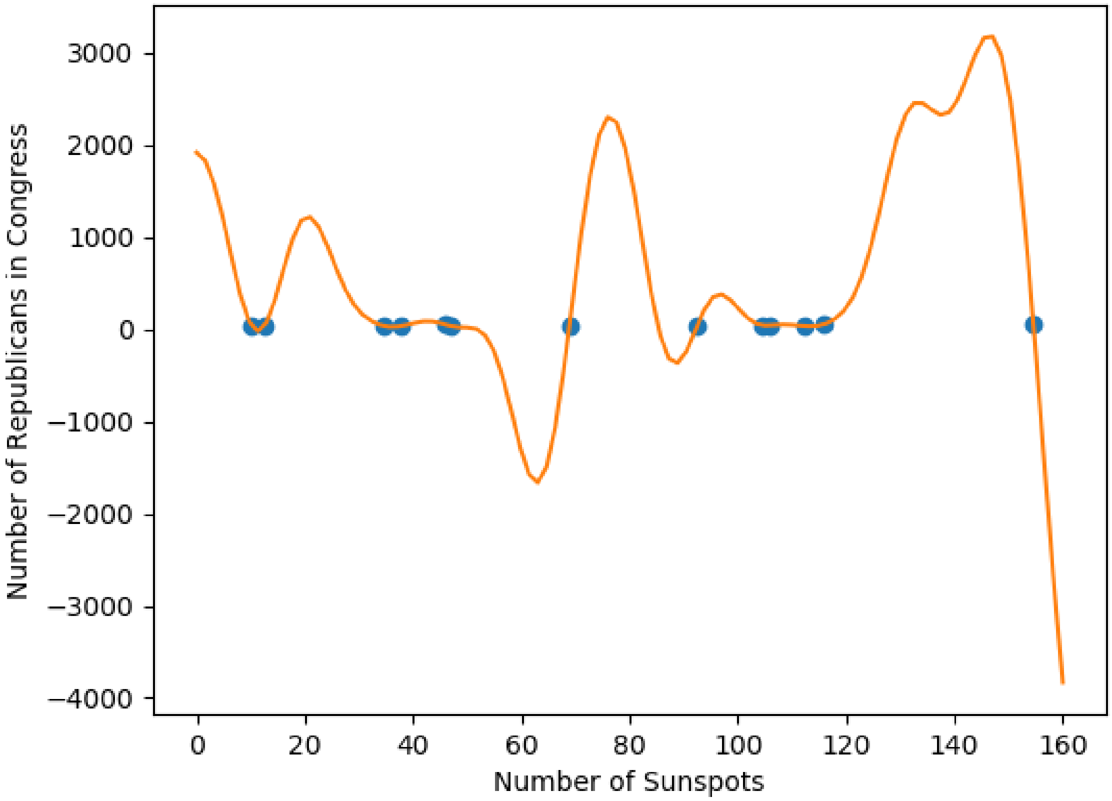
(d) This regression has a loss function value of 19.429, capturing the data patterns well. While it is numerically the best-performing model, it may be overfitting to a greater degree than model (b); the line increases strangely in the beginning and in the end, counter to the actual general data pattern.



(a) Moving on to part 3, the polynomial basis model has loss function value of 175.614. Visually, the data lacks any clear structure, so the model noticeably underfits, but it does manage to capture what appears to be a slight upward trend in the data.



(b) This basis gives a loss function value of 187.553, indicating another poor fit. Visually, the model does not attempt to capture any semblance of a trend in the data.



(c) This is the ”perfect fit”, giving a loss function value of zero. Due to the fact that there are more weights than data points, the regression line is able to pass through every data point and thus heavily overfits, failing to reflect any trend or general structure to the data. Thus, we conclude that sunspots are negligibly correlated with the number of Republicans in the Senate.

Problem 4 (Administrative)

- Name: Andrew Lu
- Email: andrewlu@college.harvard.edu
- Collaborators: Kevin Rao, Caleb He, Junu Lee
- Approximately how long did this homework take you to complete (in hours): 11