Carlos Peña-Lobel
carlospenalobel@college.harvard.edu
CS181-S16

Assignment #1
Due: 5:00pm February 5, 2016

Collaborators: Christian Hallas (kinda)

Homework 1: Linear Regression

You should submit your answers as a PDF via the Canvas course website. There is a mathematical component and a programming component to this homework.You may collaborate with others, but are expected to list collaborators, and write up your problem sets individually.

Please type your solutions after the corresponding problems using this LaTeX template, and start each problem on a new page.

Problem 1 (Centering and Ridge Regression, 7pts)
Consider a data set in which each data input vector $\boldsymbol{x} \in \mathbb{R}^m$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ be the input matrix, the rows of which are the input vectors, and the columns of which are centered at 0. Let λ be a positive constant. We define:

$$J(\boldsymbol{w}, w_0) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1}) + \lambda \boldsymbol{w}^T\boldsymbol{w}$$

(a) Compute the gradient of $J(\boldsymbol{w}, w_0)$ with respect to w_0. Simplify as much as you can for full credit.

(b) Compute the gradient of $J(\boldsymbol{w}, w_0)$ with respect to $\boldsymbol{w}$. Simplify as much as you can for full credit. Make sure to give your answer in matrix form.

(c) Suppose that $\lambda > 0$. Knowing that J is a convex function of its arguments, conclude that a global optimizer of $J(\boldsymbol{w}, w_0)$ is

$$w_0 = \frac{1}{n}\sum_i y_i \tag{1}$$

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2}$$

Before taking the inverse of a matrix, prove that it is invertible.

Solution

First we simplify:

$$J(\boldsymbol{w}, w_0) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1}) + \lambda \boldsymbol{w}^T\boldsymbol{w} \tag{3}$$

$$J(\boldsymbol{w}, w_0) = (\boldsymbol{y}^T - \boldsymbol{w}^T\boldsymbol{X}^T - \mathbf{1}^T w_0)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1}) + \lambda \boldsymbol{w}^T\boldsymbol{w} \tag{4}$$

$$J(\boldsymbol{w}, w_0) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}^T\mathbf{1}w_0 - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\mathbf{1}w_0 - \mathbf{1}^T w_0\boldsymbol{y} + \mathbf{1}^T w_0\boldsymbol{X}\boldsymbol{w} + \mathbf{1}^T\mathbf{1}w_0^2 + \lambda \boldsymbol{w}^T\boldsymbol{w} \tag{5}$$

(a) Compute the gradient of $J(\boldsymbol{w}, w_0)$ with respect to w_0. Simplify as much as you can for full credit.

Using the expanded form we take the derivative with respect to w_0

$$\bigtriangledown_{w_0} = -\boldsymbol{y}^T\mathbf{1} + \boldsymbol{w}^T\boldsymbol{X}^T\mathbf{1} - \boldsymbol{y}^T\mathbf{1} + \boldsymbol{w}^T\boldsymbol{X}^T\mathbf{1} + 2(\mathbf{1}^T\mathbf{1})w_0 \tag{6}$$

Now we can simplify $\mathbf{1}^T\mathbf{1} = n$ because this is the dot product of two 1-vectors of length n. Furthermore, because the columns of our matrix are centered at 0, $\boldsymbol{w}^T\boldsymbol{X}^T\mathbf{1} = 0$, and thus our final answer is:

$$\bigtriangledown_{w_0} = 2(nw_0 - \boldsymbol{y}^T\mathbf{1}) \tag{7}$$

(b) Compute the gradient of $J(\boldsymbol{w}, w_0)$ with respect to $\boldsymbol{w}$. Simplify as much as you can for full credit. Make sure to give your answer in matrix form.

Using the expanded form we take the derivative with respect to $\boldsymbol{w}$

$$\bigtriangledown_{\boldsymbol{w}} = -\boldsymbol{y}^T\boldsymbol{X} - \boldsymbol{y}^T\boldsymbol{X} + 2(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}) + \mathbf{1}^T\boldsymbol{X}w_0 + \mathbf{1}^T\boldsymbol{X}w_0 + 2\lambda\boldsymbol{w}^T \tag{8}$$

Once again, because the columns of our matrix are centered at 0, $\mathbf{1}^T\boldsymbol{X}w_0 = 0$, and thus our final answer is:

$$\bigtriangledown_{\boldsymbol{w}} = 2(\lambda\boldsymbol{w}^T - \boldsymbol{y}^T\boldsymbol{X} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}) \tag{9}$$

(c) Suppose that $\lambda > 0$. Knowing that J is a convex function of its arguments, conclude that a global optimizer of $J(\boldsymbol{w}, w_0)$ is

$$w_0 = \frac{1}{n}\sum_i y_i \tag{10}$$

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{11}$$

Before taking the inverse of a matrix, prove that it is invertible.

Since J is convex, we only need to see where the gradient is 0. Thus we set our values in lines (7) and (9) equal to 0:

$$0 = 2(nw_0 - \boldsymbol{y}^T\mathbf{1}) \tag{12}$$

$$nw_0 = \boldsymbol{y}^T\mathbf{1} \tag{13}$$

$$w_0 = \frac{1}{n}\sum_i y_i \tag{14}$$

$$\tag{15}$$

$$0 = 2(\lambda\boldsymbol{w}^T - \boldsymbol{y}^T\boldsymbol{X} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}) \tag{16}$$

$$\left[\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{w}^T\right]^T = \left[\boldsymbol{y}^T\boldsymbol{X}\right]^T \tag{17}$$

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y} \tag{18}$$

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y} \tag{19}$$

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{20}$$

Now to prove that the matrix is invertible, we will show that the matrix is Positive-definite. A positive-definite matrix $\boldsymbol{A} \implies$ eigenvalues are positive, $\implies det\boldsymbol{A} > 0$, $\implies \boldsymbol{A}$ is invertible.

$$\begin{aligned} \text{Positive-definite matrix } \boldsymbol{A} &\implies \text{positive eigenvalues,} \\ &\implies \det \boldsymbol{A} > 0 \\ &\implies \boldsymbol{A} \text{ is invertible.} \end{aligned}$$

Now to show that a matrix is positive-definite we must show that

$$\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} > 0 \ \forall \boldsymbol{v} \neq 0$$

.

Looking at our matrix $\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}$:

$$\boldsymbol{v}^T (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}) \boldsymbol{v} > 0 \tag{21}$$

$$\boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{v} + \boldsymbol{v}^T \lambda \boldsymbol{I} \boldsymbol{v} > 0 \tag{22}$$

$$\|\boldsymbol{X}\boldsymbol{v}\| + \lambda \|\boldsymbol{v}\| > 0 \tag{23}$$

Since these are both lengths, we know that they must both be greater than or equal to 0 in all cases. Moreover, since the definition defines $v \neq 0$, we know that the final term must be greater than 0 proving that the matrix is in fact positive definite and thus invertible.

Problem 2 (Priors and Regularization,7pts)
Consider the Bayesian linear regression model given in Bishop 3.3.1. The prior is

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\boldsymbol{I}),$$

where α is the precision parameter that controls the variance of the Gaussian prior. The likelihood can be written as

$$p(\boldsymbol{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}),$$

Using the fact that the posterior is the product of the prior and the likelihood (up to a normalization constant), show that maximizing the log posterior (i.e., $\ln p(\boldsymbol{w} \mid \boldsymbol{t}) = \ln p(\boldsymbol{w}|\alpha) + \ln p(\boldsymbol{t} \mid \boldsymbol{w})$) is equivalent to minimizing the regularized error term given by $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$ with

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}_n))^2$$
$$E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}$$

Do this by writing $\ln p(\boldsymbol{w} \mid \boldsymbol{t})$ as a function of $E_D(\boldsymbol{w})$ and $E_W(\boldsymbol{w})$, dropping constant terms if necessary. Conclude that maximizing this posterior is equivalent to minimizing the regularized error term given by $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$. (Hint: take $\lambda = \alpha/\beta$)

Solution

First let us consider the log-likelihood for a standard normal, $\mathcal{N}(x \mid \mu, \sigma^2)$.

$$p(x \mid \mu, \sigma^2) = \prod^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$
$$\ln\left(p(x \mid \mu, \sigma^2)\right) = \sum^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum^{N} -\frac{1}{2\sigma^2}(x-\mu)^2$$
$$\ln\left(p(x \mid \mu, \sigma^2)\right) = N\left(\ln\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\right) - \frac{1}{2\sigma^2}\sum_{i}^{N}(x_i-\mu)^2$$
$$\ln\left(p(x \mid \mu, \sigma^2)\right) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i}^{N}(x_i-\mu)^2$$

Since both of our values are normally distributed, we can throw them into this key, dropping additive constants and adding them together (because we have taken the log, multiplicative constants are now additive). This yields:

$$\ln(p(\boldsymbol{w} \mid \alpha) * p(\boldsymbol{t} \mid \boldsymbol{w})) = \ln(\mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) + \ln \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1})$$

$$\ln(p(\boldsymbol{w} \mid \alpha) * p(\boldsymbol{t} \mid \boldsymbol{w})) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln\left(\alpha^{-1}\boldsymbol{I}\right) - \frac{\alpha}{2\boldsymbol{I}} \sum_{i}^{N} (w_i - 0)^2 - \frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln\left(\beta^{-1}\right) - \frac{\beta}{2} \sum_{n}^{N} \left(t_n - \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_n)\right)^2$$

$$\ln(p(\boldsymbol{w} \mid \alpha) * p(\boldsymbol{t} \mid \boldsymbol{w})) = -\left(\frac{\alpha}{2\boldsymbol{I}} \sum_{i}^{N} (w_i)^2 + \frac{\beta}{2} \sum_{n}^{N} \left(t_n - \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_n)\right)^2\right)$$

$$\ln(p(\boldsymbol{w} \mid \alpha) * p(\boldsymbol{t} \mid \boldsymbol{w})) = -\left(\frac{\alpha}{2\beta} \boldsymbol{w}^T \boldsymbol{w} + \frac{1}{2} \sum_{n}^{N} \left(t_n - \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_n)\right)^2\right)$$

$$\ln(p(\boldsymbol{w} \mid \alpha) * p(\boldsymbol{t} \mid \boldsymbol{w})) = -\left(\lambda E_W(\boldsymbol{w}) + E_D(\boldsymbol{w})\right) \text{ If } \lambda = \frac{\alpha}{\beta}$$

Since this is negative, and we are maximizing this value, we clearly want to minimize the positive of this value.

3. Modeling Changes in Congress [10pts]

The objective of this problem is to learn about linear regression with basis functions by modeling the average age of the US Congress. The file `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. Currently, the 114th Congress is in session. The second is the average age of that members of that Congress. The data file looks 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 in Figure 1.

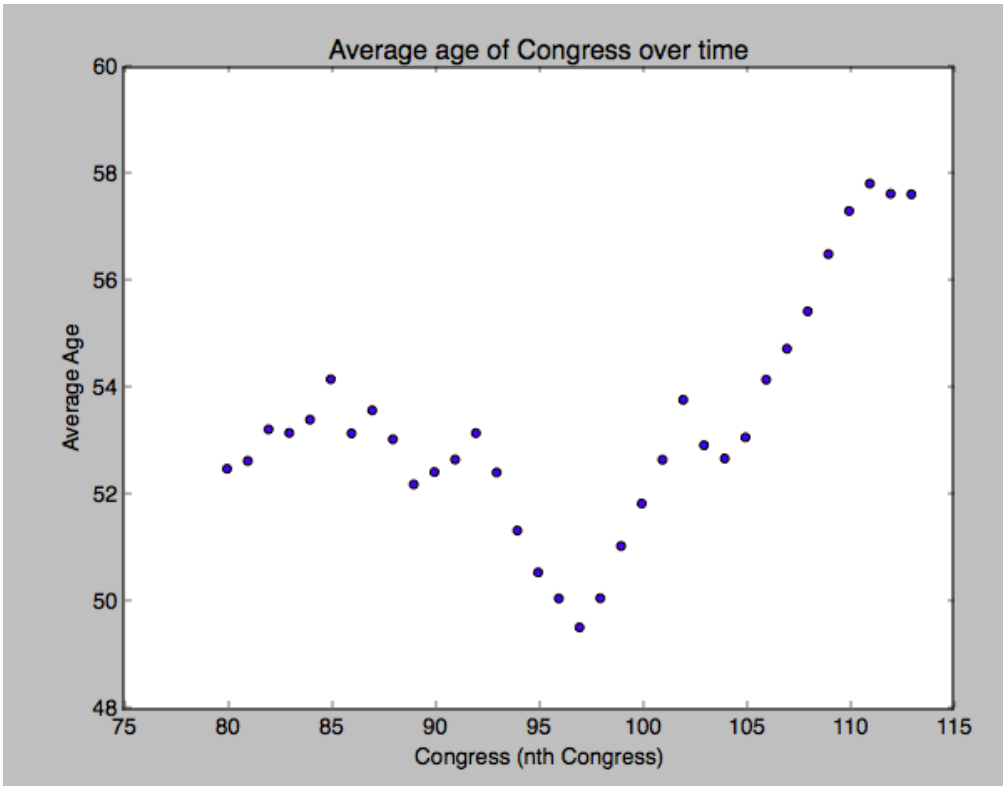

Figure 1: Average age of Congress. The horizontal axis is the Congress number, and the vertical axis is the average age of the congressmen.

Problem 3 (Modeling Changes in Congress, 10pts)

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 $j = 1, \ldots, 7$

(b) $\phi_j(x) = x^j$ for $j = 1, \ldots, 3$

(c) $\phi_j(x) = \sin\{x/j\}$ for $j = 1, \ldots, 4$

(d) $\phi_j(x) = \sin\{x/j\}$ for $j = 1, \ldots, 7$

(e) $\phi_j(x) = \sin\{x/j\}$ for $j = 1, \ldots, 20$

In addition to the plots, provide one or two sentences for each, 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.

Solution

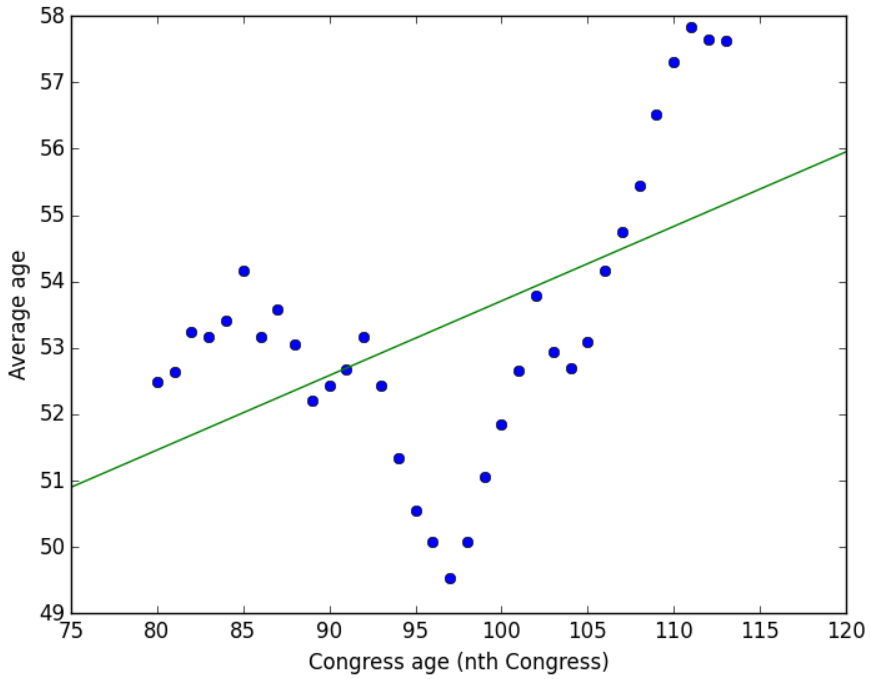

Figure 2: This is a simple linear regression, and clearly looks to be underfitting because it is missing the whole downward spike in the middle.

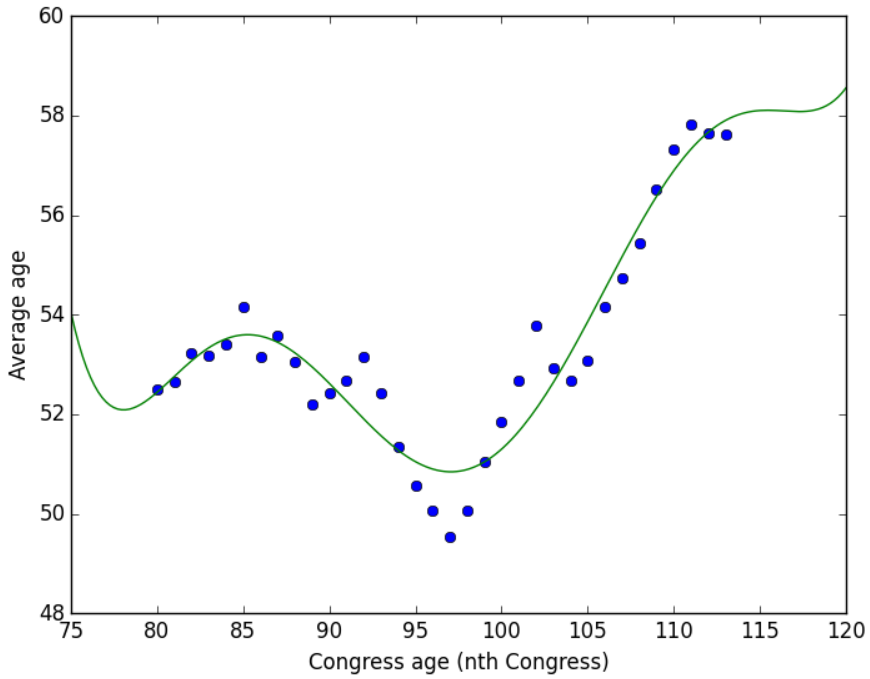

Figure 3: This example seems to be fitting the data better at the trend of a dropping of the congressional age. However since this uses terms to the 7th power, it may not be good at predicting future ages of Congress.

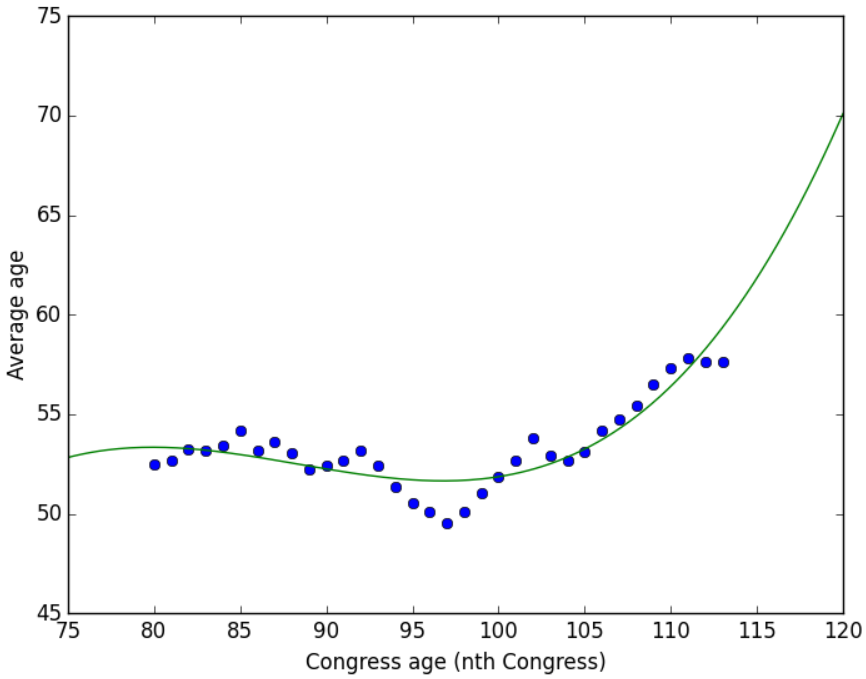

Figure 4: While this generally fits the previous data, the fit grows exponentially in the future, which likely won't model future Congressional compositions.

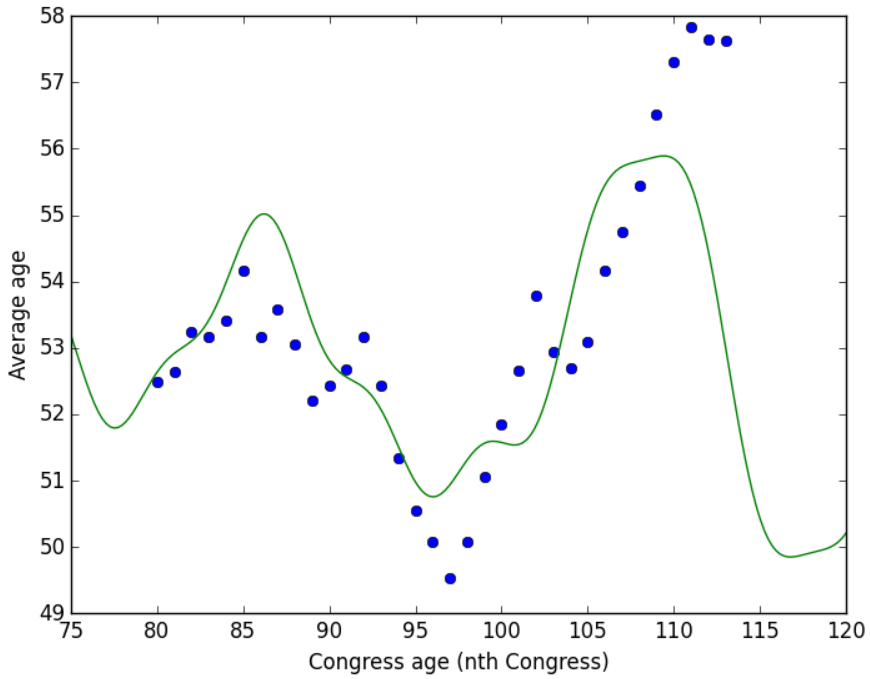

Figure 5: This is underfitting the data, missing the drop of age around the 97th congress and overinflating a spike after the 86th.

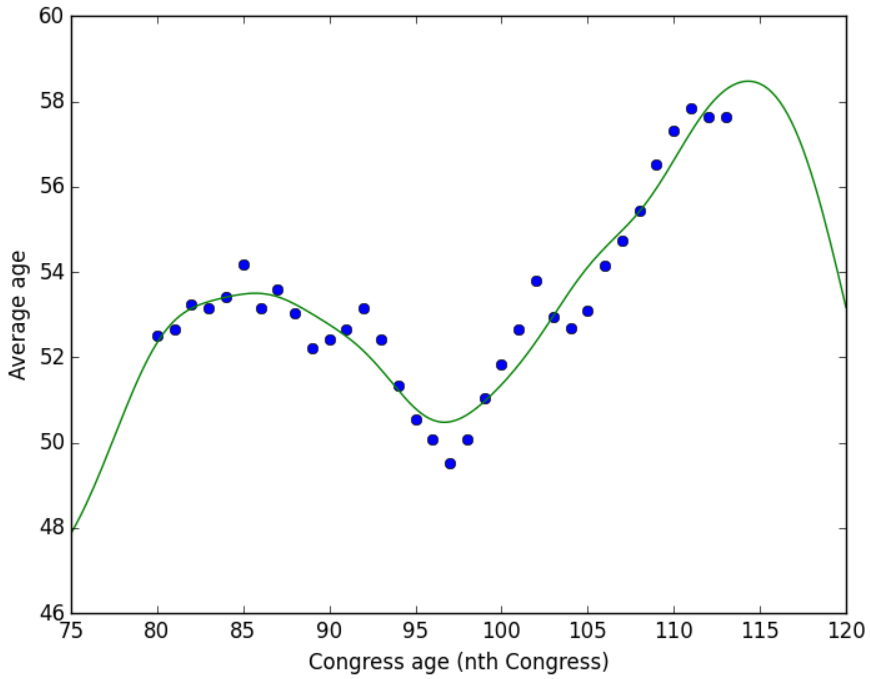

Figure 6: This fits the data very well, not overvaluing any parts. A word of caution is that it quickly expects the age of Congress to decrease in the future.

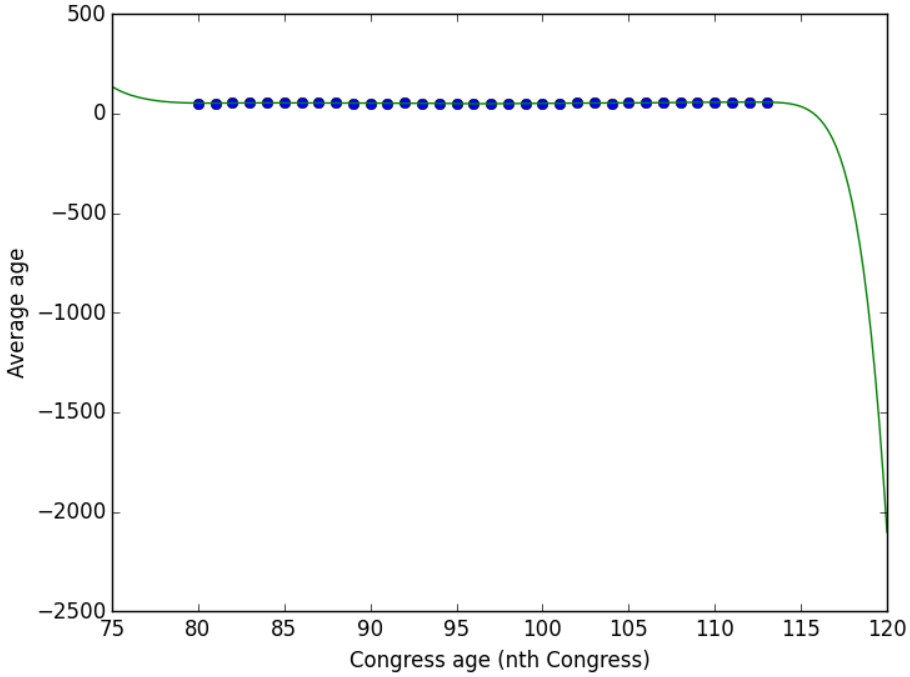

Figure 7: This figure super overfits the data. The fact that there are essentially as many terms as datapoints allows the program to draw the line directly through every point. This has 0 predictive value because this model is so overfit to past data, it doesn't try to distinguish the signal from the noise.

Problem 4 (Calibration, 1pt)
Approximately how long did this homework take you to complete?

Answer: 12