Matt Leifer
matthewleifer@college.harvard.edu
CS181-S17

Assignment #1
Due: 5:00pm February 3, 2017

Collaborators: Tomislav Zabcic-Matic

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.

Problem 1 (Centering and Ridge Regression, 7pts)

Consider a data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ in which each input vector $\mathbf{x} \in \mathbb{R}^m$. As we saw in lecture, this data set can be written using the design matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ and the target vector $\mathbf{y} \in \mathbb{R}^n$.

For this problem assume that the input matrix is centered, that is the data has been pre-processed such that $\frac{1}{n}\sum_{i=1}^n x_{ij} = 0$. Additionally we will use a positive regularization constant $\lambda > 0$ to add a ridge regression term.

In particular we consider a ridge regression loss function of the following form,

$$\mathcal{L}(\mathbf{w}, w_0) = (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})^\top(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + \lambda\mathbf{w}^\top\mathbf{w}.$$

Note that we are not incorporating the bias $w_0 \in \mathbb{R}$ into the weight parameter $\mathbf{w} \in \mathbb{R}^m$. For this problem the notation $\mathbf{1}$ indicates a vector of all 1's, in this case in implied to be in $\mathbb{R}^n$.

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

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

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

$$w_0 = \frac{1}{n}\sum_{i=1}^n y_i \tag{1}$$

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

(d) In order to take the inverse in the previous question, the matrix $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$ must be invertible. One way to ensure invertibility is by showing that a matrix is *positive definite*, i.e. it has all positive eigenvalues. Given that $\mathbf{X}^\top\mathbf{X}$ is positive *semi*-definite, i.e. all non-negative eigenvalues, prove that the full matrix is invertible.

(e) What difference does the last problem highlight standard least-squares regression versus ridge regression?

Solution

1. $\frac{\mathrm{d}}{\mathrm{d}w_0}\mathcal{L}(\mathbf{w}) = (\frac{\mathrm{d}}{\mathrm{d}w_0}(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}))^T((\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})) + (\frac{\mathrm{d}}{\mathrm{d}w_0}(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}))^T((\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})) + 0$
   $= -2 \cdot \mathbf{1}^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})$
   $$= -2\sum_{i=1}^n (y_i - w_0 - \sum_{j=1}^m w_j x_{ij})$$
   $$= 2nw_0 - 2\sum_{i=1}^n y_i - 2\sum_{j=1}^m w_j \sum_{i=1}^n x_{ij}$$
   $$= 2nw_0 - 2\sum_{i=1}^n y_i \text{ because } \sum_{i=1}^n x_{ij} = 0.$$

2. $\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\mathcal{L}(\mathbf{w}) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + 2\lambda\mathbf{w}$ by using the rule that $\frac{d}{d\mathbf{x}}\mathbf{x}^T\mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$, where in this case $\mathbf{A} = \mathbf{I}$, the identity matrix.

3. Set $\frac{d}{dw_0}\mathcal{L}(\mathbf{w})$ to 0.
   $$0 = 2nw_0 - 2\sum_{i=1}^n y_i$$

$nw_0 = \sum_{i=1}^{n} y_i$

$w_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ which is what we wanted to show.

Set $\frac{d}{d\mathbf{w}}\mathcal{L}(\mathbf{w})$ to 0.
$0 = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + 2\lambda\mathbf{w}$
$\lambda\mathbf{w} = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T w_0\mathbf{1}$
$\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T w_0\mathbf{1}$
$(\mathbf{X}^T\mathbf{X} + +\lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T w_0\mathbf{1}$

However, $\mathbf{X}^T w_0\mathbf{1} = 0$ because $\mathbf{X}^T w_0\mathbf{1} = w_0 \cdot [\sum_{i=1}^{n} x_{i1} \sum_{i=1}^{n} x_{i2} ... \sum_{i=1}^{n} x_{im}]^T$ and we know that since the data is centered that $\sum_{i=1}^{n} x_{ij} = 0$
So, $(\mathbf{X}^T\mathbf{X} + +\lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{y}$
$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + +\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ which is what we wanted to show.
Because we're taking that $\mathcal{L}(\mathbf{w}, w_0)$ is convex in its arguments and at these values of $\mathbf{w}$ and w_0 the derivatives of $\mathcal{L}$ are 0, we know that these values are the global optimizers.

4. Let e be an eigenvalue of $\mathbf{X}^T\mathbf{X}$. Because $\mathbf{X}^T\mathbf{X}$ is semi-definite, $e \geq 0$. $(\mathbf{X}^T\mathbf{X} - e\mathbf{I})\mathbf{v} = 0$.
Now consider the eigenvalues of $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$.
$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{v} = k\mathbf{v}$
$(\mathbf{X}^T\mathbf{X} + (\lambda - k)\mathbf{I})\mathbf{v} = 0$
Therefore $\lambda - k = -e$ and so $k = \lambda - e$. Therefore because $\lambda > 0$ and $e \geq 0$, $k > 0$. That is to say that all the eigenvalues of $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ are greater than zero. This implies that $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is invertible.

5. In least squares regression, we have to take the inverse of $\mathbf{X}^T\mathbf{X}$ and because the eigenvalues of this can be 0, this matrix may not be invertible. However, the matrix in ridge regression is always invertible.

Problem 2 (Priors and Regularization,7pts)

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

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

where α is as scalar precision hyperparameter that controls the variance of the Gaussian prior. Define the likelihood as,

$$p(\mathbf{y} \,|\, \mathbf{x}) = \prod_{i=1}^{n} \mathcal{N}(y_i \,|\, \mathbf{w}^\mathsf{T}\mathbf{x}_i, \beta^{-1}),$$

where β is another fixed scalar defining the variance.

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

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

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

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

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

Solution

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

$p(\mathbf{w}) = \frac{1}{(2\pi)^{0.5D}} \cdot \frac{1}{(\det(\alpha^{-1}\mathbf{I}))^{0.5}} \exp(-\frac{1}{2}\mathbf{w}^T\alpha\mathbf{I}\mathbf{w})$

$\ln p(\mathbf{w}) = c_1 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} = -\alpha\mathcal{R}(\mathbf{w}) + c_1$

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{n}\mathcal{N}(y_i|\mathbf{w}^T x_i, \beta^-1)$$
$$p(\mathbf{y}|\mathbf{w}) = \sum_{i=1}^{n}\ln(\frac{1}{(2\pi\beta^{-1})^{0.5}}) - \frac{1}{2\beta^{-1}}(y_i - \mathbf{w}^T\mathbf{x})^2 = c_2 - \beta\mathcal{L}(\mathbf{w})$$

Therefore, $\ln p(\mathbf{w}|\mathbf{y}) = -\alpha\mathcal{R}(\mathbf{w}) - \beta\mathcal{L}(\mathbf{w}) + c_1 + c_2$ up to a normalization class. If we drop $c_1 + c_2$, which is a constant then we can see that $\ln(p(\mathbf{w}|\mathbf{y}))$ is proportional to $-(\frac{\alpha}{\beta}\mathcal{R}(\mathbf{w}) + \mathcal{L}(\mathbf{w})$.

We know that $\mathcal{R}(\mathbf{w})$ and $\mathcal{L}(\mathbf{w})$ are both always greater than or equal to zero ($\mathcal{L}$ is the sum of squared terms and $\mathcal{R}$ is equivalent to half the dot product of $\mathbf{w}$) so if we minimize this error term ($\mathcal{L}(\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w})$, where $\lambda = \frac{\alpha}{\beta}$) we can maximize the log of the posterior probability and by extension of the posterior probability.

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.

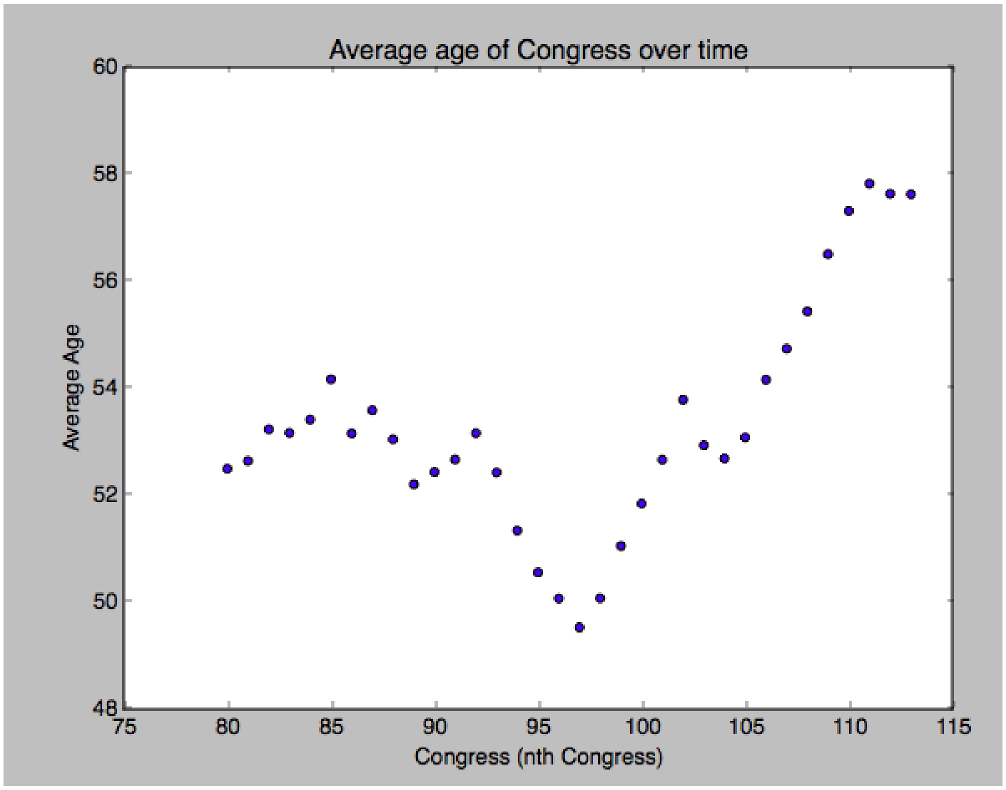


Figure 1: Average age of Congress. The horizontal axis is the Congress number, and the vertical axis is the average age of the congressmen.

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 $j = 1, \ldots, 6$

(b) $\phi_j(x) = x^j$ for $j = 1, \ldots, 4$

(c) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 6$

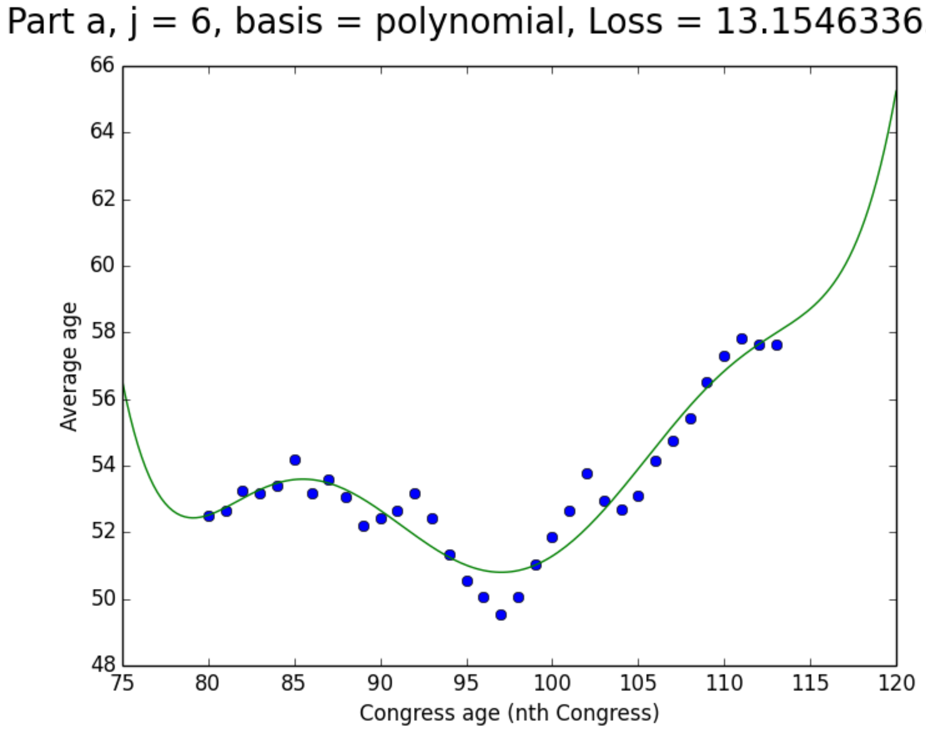
(d) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 10$

(e) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 22$

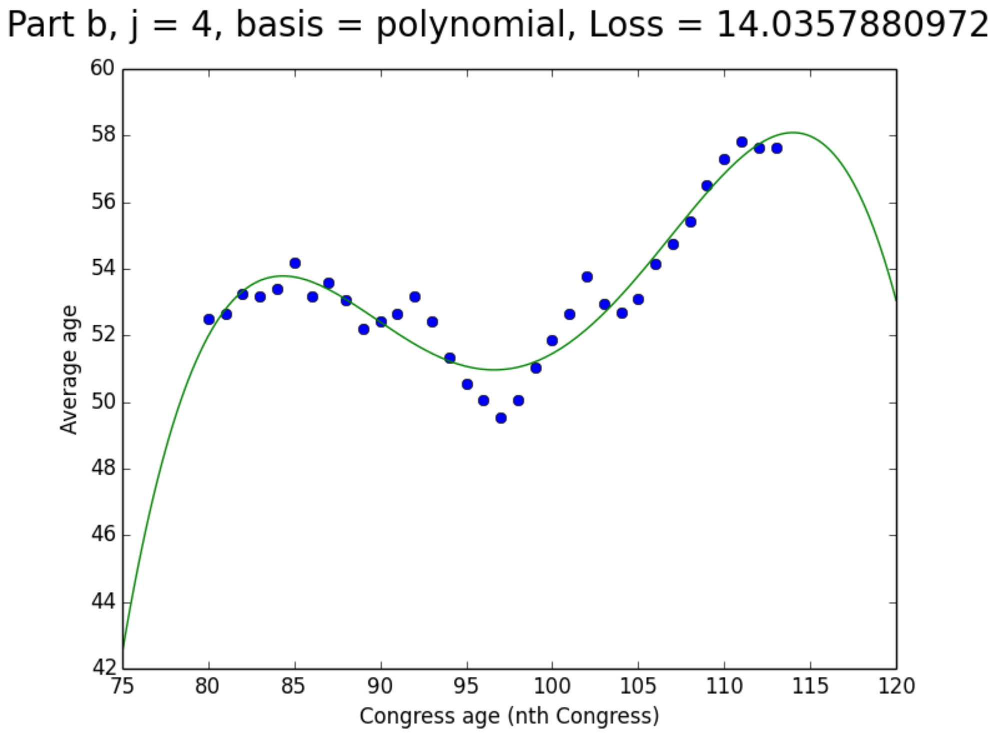
In addition to the plots, provide one or two sentences for each, explaining whether you think it is fitting well, overfitting or underfitting. If it does not fit well, provide a sentence explaining why. A good fit should capture the most important trends in the data.

Solution

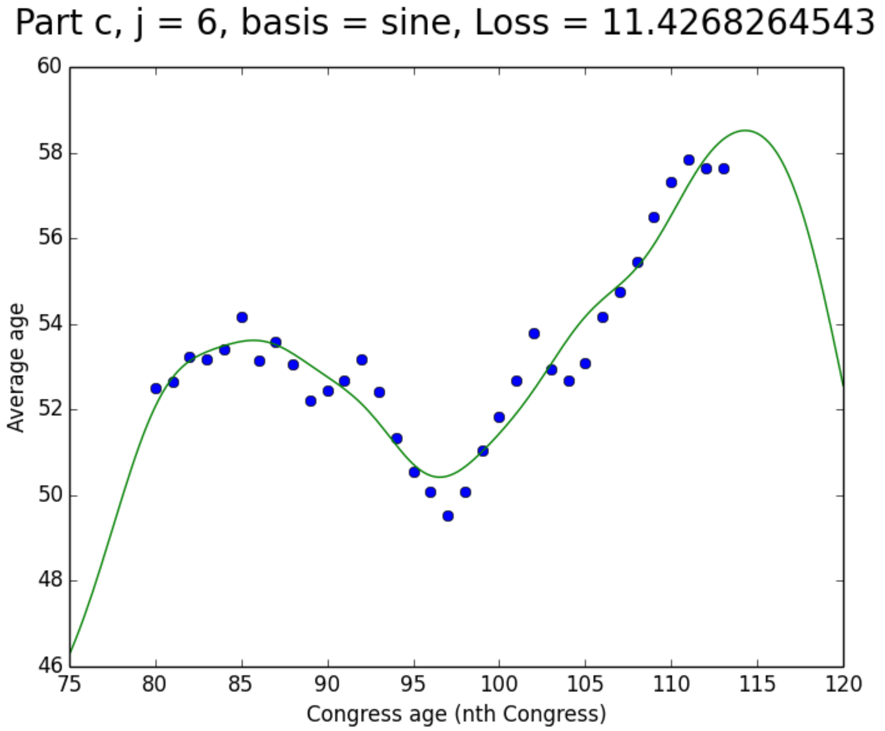
1. The "Loss" calculated in the titles of all these graphs is the sum of the squares of the residuals. It is the same as the $\mathcal{L}$ function in question 2.



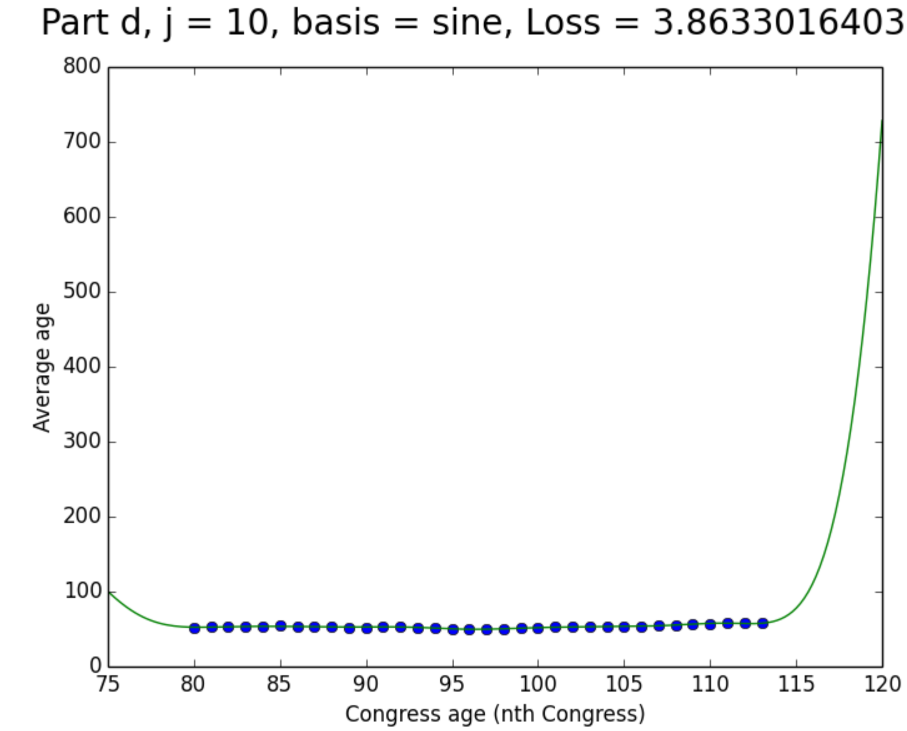
This graph isn't bad but it slightly overfits the data. As soon as we get out of the range of the data set the values of average age grow extremely quickly.



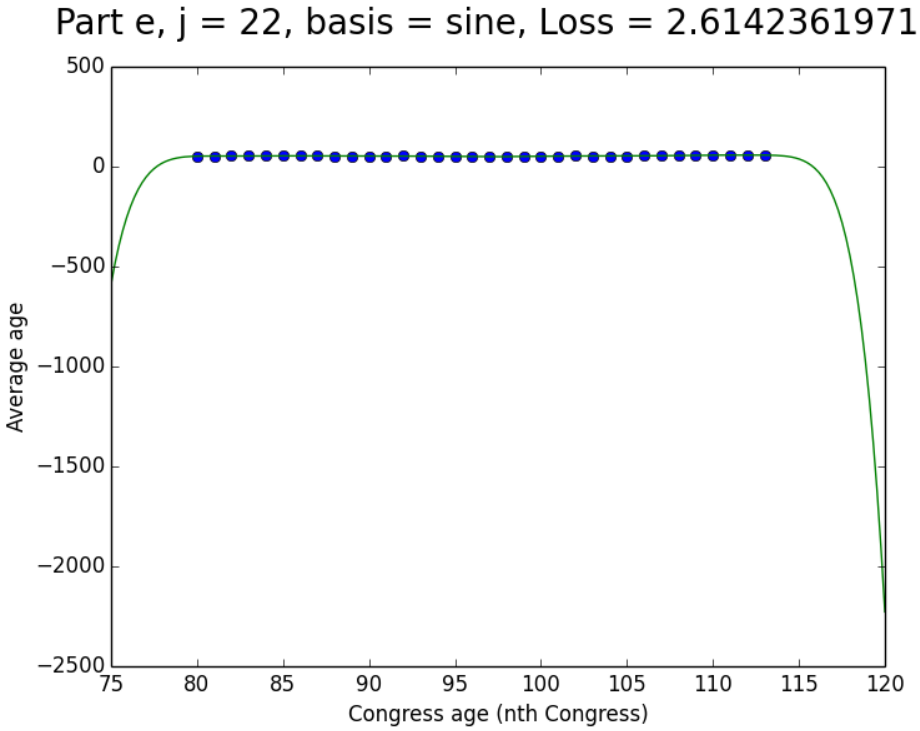
This one captures the overall shape of the data and might be good for a few years outside of the range of data but since it's a polynomial eventually the x^4 term will dominate and the data will go to infinity - the underlying basis function don't model the age of people accurately.



This graph probably fits the data the best of all the ones here, it's close to the data and captures the kinks in the data and also looks like it gives reasonable predicitons out side the range.



This is definitely a case of overfitting the data. The curve hits the given data extremely well and is the 2nd best of these graphs to minimize the error but almost as soon as we leave the domain of the data the prediction is wildly off.



Just like the graph before, this overfits the data and probably uses too many parameters. So the error is very low but the function is very bad for predictions.

Problem 4 (Calibration, 1pt)

Approximately how long did this homework take you to complete?

Answer: 8hrs