Repo: cs181-s17-hixor
Hikari Sorensen
laurensorensen@college.harvard.edu
CS181-S17

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

Collaborators: None

Homework 1: Linear Regression

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

(a)

$$\frac{\partial\mathcal{L}}{\partial w_0} = -2(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})$$

since, when taking derivatives, we can treat $M^\top M$ as M^2 for some matrix M. This is clear if we first expand into summation notation, then put back into matrix form. Then, we can just treat this like any other partial.

(b)

$$\frac{\partial\mathcal{L}}{\partial \mathbf{w}} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + 2\lambda\mathbf{w}$$

Similar to in (a), we treat $M^\top M$ for some matrix Mas M^2, then take the partial. The chain rule yields a factor of $\mathbf{X}^\top$ in the first term, which we transpose in order to keep matrix shapes compatible.

(c) Since $\mathcal{L}$ is convex, we know that there exists a unique global minimum. We can find this minimum by setting $\nabla\mathcal{L} = 0$.
First, set $\frac{\partial\mathcal{L}}{\partial w_0} = 0$.

$$0 = \frac{\partial\mathcal{L}}{\partial w_0} = -2(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})$$

$$
\begin{aligned}
0 = \frac{\partial \mathcal{L}}{\partial \mathbf{w}} &= -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + 2\lambda\mathbf{w} \\
\implies \quad 0 &= \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) - \lambda\mathbf{w} \\
\implies \quad 0 &= \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top w_0\mathbf{1} - \lambda\mathbf{w} \\
\implies \quad \lambda\mathbf{w} + \mathbf{X}^\top\mathbf{X}\mathbf{w} &= \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top w_0\mathbf{1} \\
\implies \quad (\lambda\mathbf{I} + \mathbf{X}^\top\mathbf{X})\mathbf{w} &= \mathbf{X}^\top\mathbf{y} \qquad (1') \\
\implies \quad \mathbf{w} &= (\lambda\mathbf{I} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}
$$

where the $(-\mathbf{X}^\top w_0\mathbf{1})$ term goes to 0 in the right side of $(1')$ because $\mathbf{X}$ is centered, and multiplying it by a constant keeps it centered (i.e., sum still equals 0).

(d) If $\mathbf{X}^\top\mathbf{X}$ is positive semi-definite, then it has all non-negative eigenvalues. Adding $\lambda\mathbf{I}$ adds λ to each of the elements along the diagonal. If $\lambda > 0$, then all eigenvalues increase in value. In particular, any 0 eigenvalues become positive.

(e) Beyond having the effect of encouraging simpler models by penalizing many and large weights, part (d) shows that ridge regression yields a matrix $M = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$ such that M is always invertible, unlike the positive semi-definite matrix $M^\top M$ of ordinary least squares, which is not necessarily invertible. Thus, ridge regression always has a unique analytical solution, whereas OLS may not (in which case, we use a different optimization technique, like gradient descent).

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

Let us first find the values for $\ln\, p(\mathbf{w})$ and $\ln\, p(\mathbf{y}|\mathbf{w})$.

$$\ln p(\mathbf{w}) = \ln\left(\frac{1}{\sqrt{2\pi\alpha^{-1}\mathbf{I}}}\exp\{-\frac{1}{2}\mathbf{w}^{\top}(\alpha^{-1}\mathbf{I})\mathbf{w}\}\right) = c - \frac{1}{2}\mathbf{w}^{\top}(\alpha^{-1}\mathbf{I})\mathbf{w} = c - \alpha^{-1}\frac{1}{2}\mathbf{w}^{\top}\mathbf{w}$$

where

$$c = \ln\left(\frac{1}{\sqrt{2\pi\alpha^{-1}\mathbf{I}}}\right).$$

Also,

$$\ln p(\mathbf{y}|\mathbf{w}) = \ln\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\beta^{-1}}}\exp\{-\frac{1}{2\beta^{-1}}(y_i - \mathbf{w}^{\top}\mathbf{x}_i)^2\}\right) = d - \frac{1}{\beta^{-1}}\frac{1}{2}\sum_{i=n}^{n}(y_i - \mathbf{w}^{\top}\mathbf{x}_i)^2$$

where

$$d = \ln\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\beta^{-1}}}\right)$$

Now, notice how if we drop the constant values c and d in $\ln p(\mathbf{w})$ and $\ln p(\mathbf{y}|\mathbf{w})$, respectively, we have that

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

$$\implies \quad \arg\max_{\mathbf{w}} \ln p(\mathbf{w}|\mathbf{y}) = \arg\max_{\mathbf{w}}(-\alpha^{-1}\frac{1}{2}\mathbf{w}^\top\mathbf{w} - \frac{1}{\beta^{-1}}\frac{1}{2}\sum_{i=n}^{n}(y_i - \mathbf{w}^\top\mathbf{x}_i)^2)$$
$$= \arg\max_{\mathbf{w}}(-\alpha^{-1}\mathcal{R}(\mathbf{w}) - \frac{1}{\beta^{-1}}\mathcal{L}(\mathbf{w}))$$
$$= \arg\min_{\mathbf{w}}(\alpha^{-1}\mathcal{R}(\mathbf{w}) + \frac{1}{\beta^{-1}}\mathcal{L}(\mathbf{w}))$$
$$= \arg\min_{\mathbf{w}}(\beta^{-1}\alpha^{-1}\mathcal{R}(\mathbf{w}) + \mathcal{L}(\mathbf{w}))$$

since we can factor out a $\frac{1}{\beta^{-1}}$ term, then drop it from the overall quantity we want to minimize, since the desired minimization is of a quadratic, and so has some global minimum that isn't affected by the coefficient in front. Thus,

$$\arg\max_{\mathbf{w}} \ln p(\mathbf{w} \mid \mathbf{y}) = \arg\min_{\mathbf{w}}(\lambda\mathcal{R}(\mathbf{w}) + \mathcal{L}(\mathbf{w}))$$

where $\lambda = \frac{1}{\alpha\beta}$, as desired. □

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.

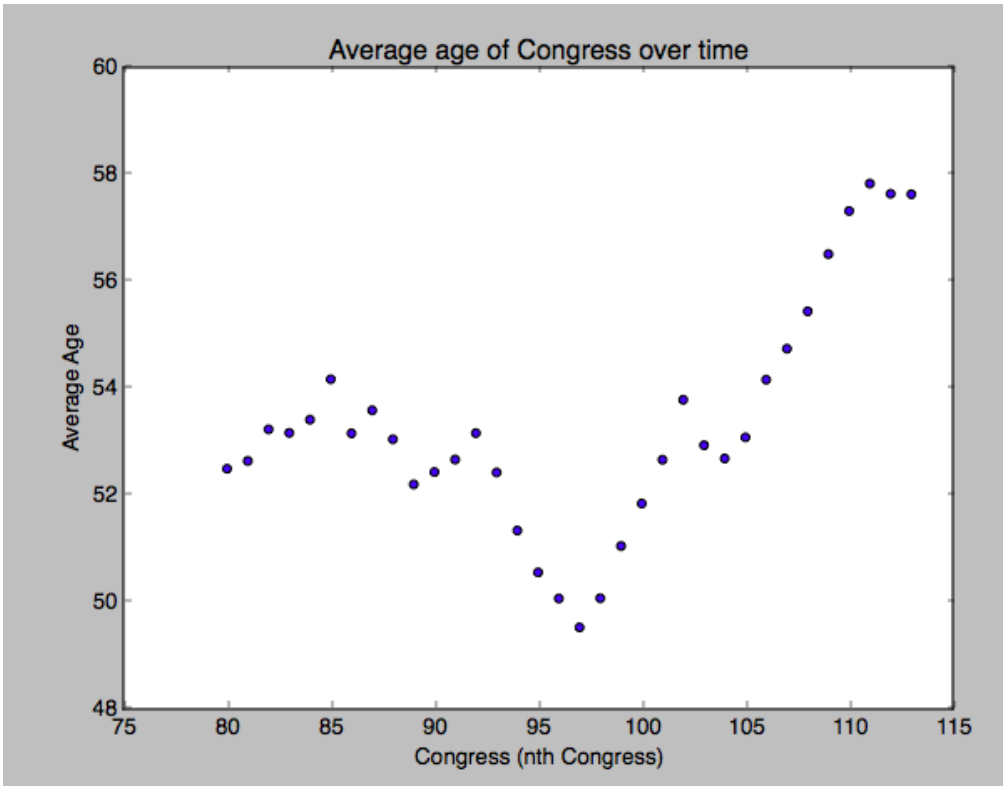


Figure 1: Average age of Congress. The horizontal axis is the Congress number, and the vertical axis is the average age of the congressmen.

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 $j = 1, \ldots, 6$

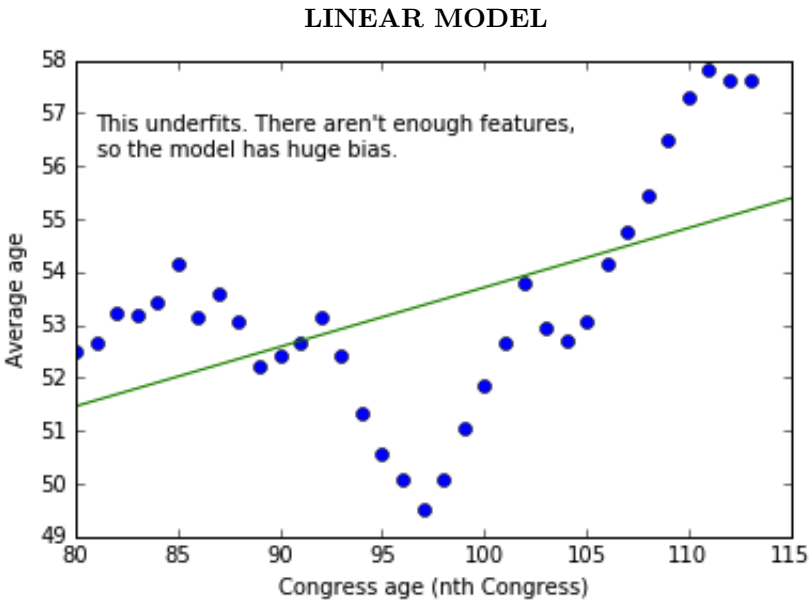
(d) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 10$

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

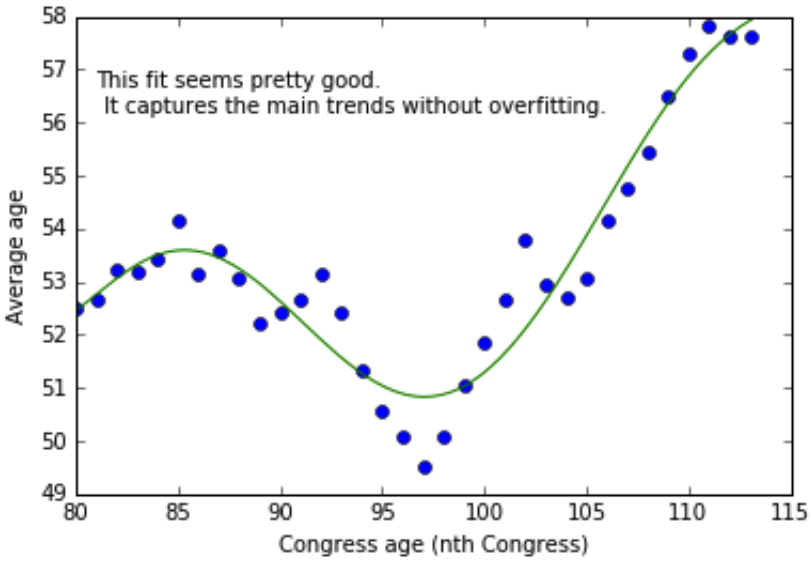
In addition to the plots, provide one or two sentences for each, explaining whether you think it is fitting well, overfitting or underfitting. If it does not fit well, provide a sentence explaining why. A good fit should capture the most important trends in the data.

Solution

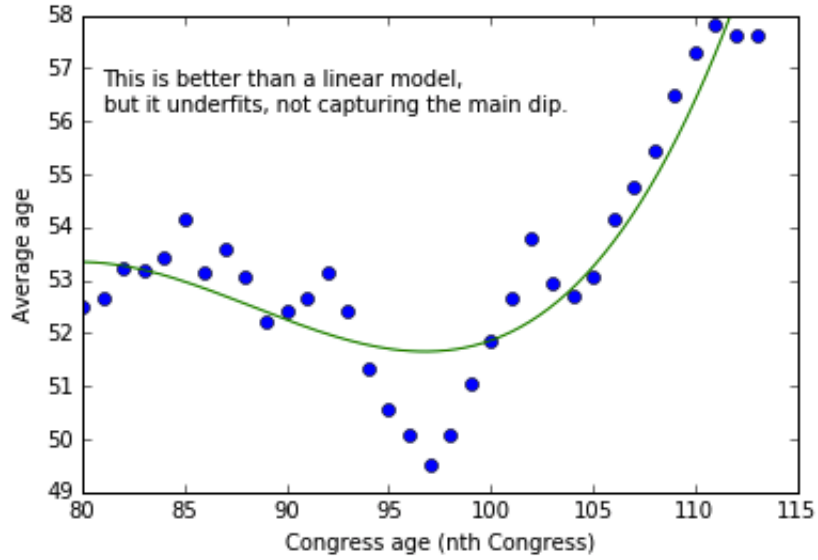
See the subitted linreg.py file for code. The graphs produced are as follows:



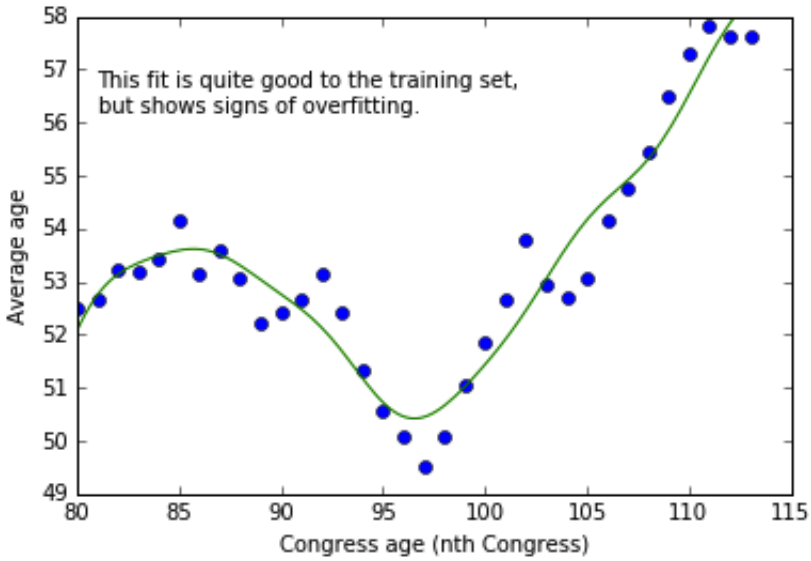
USING BASIS (a)



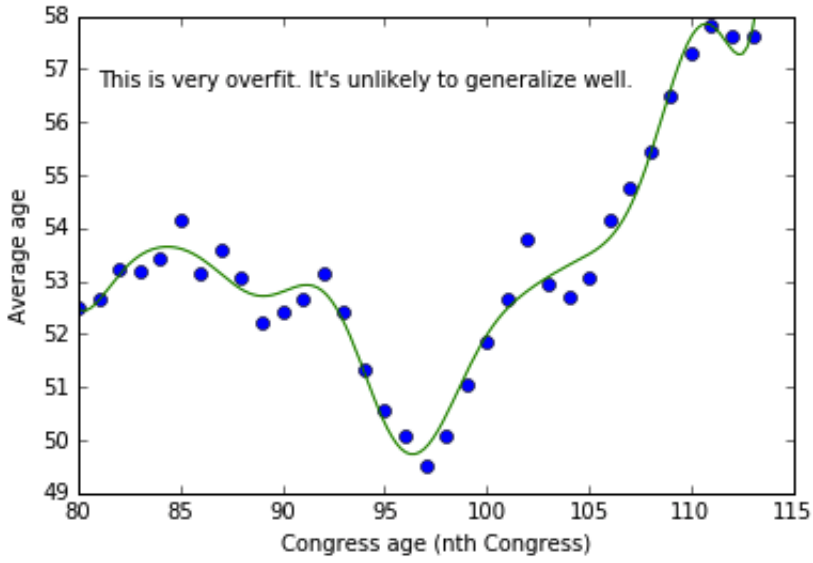
USING BASIS (b)



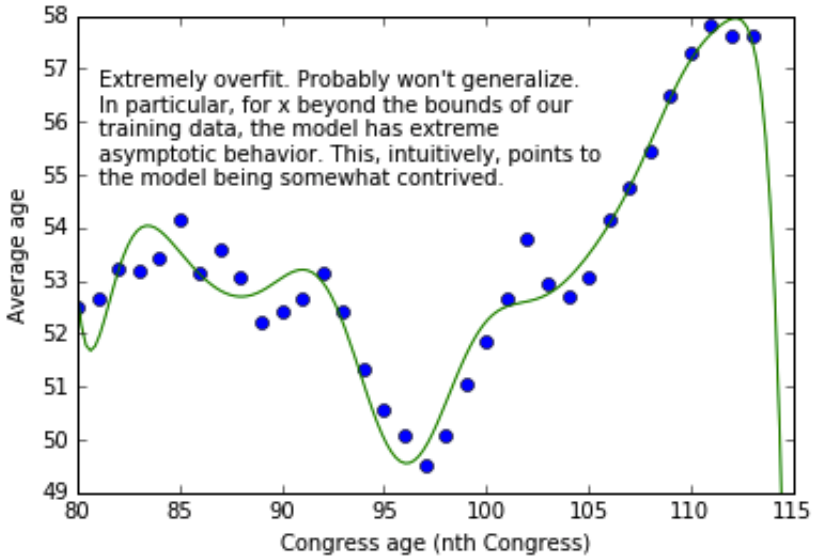
USING BASIS (c)



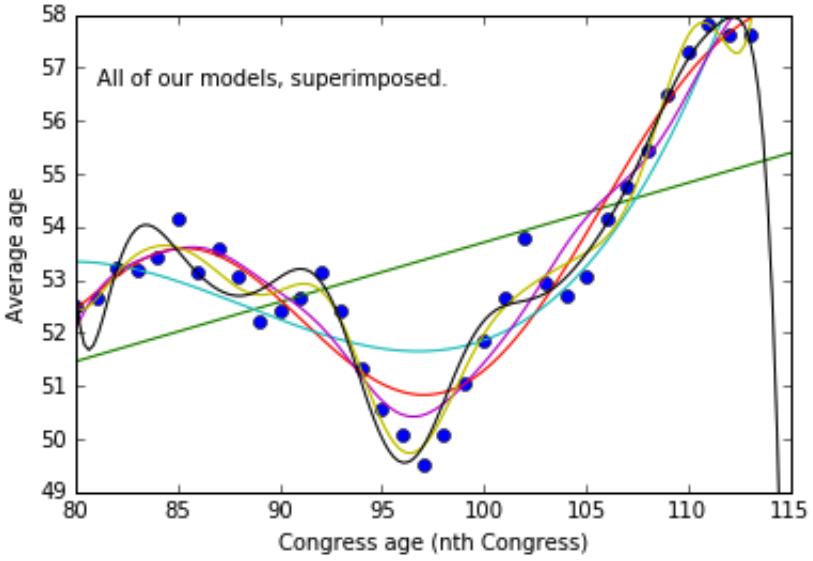
USING BASIS (d)



USING BASIS (e)



ALL BASES, SUPERIMPOSED



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

Answer: Problems themselves: on the order of 2-3 hours. Just LaTex-ing the entire thing, and making everything look nice: on the order of 5 hours.