Matheus C. Fernandes
fernandes@fas.harvard.edu
CS181-S17

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

Collaborators:

Homework 1: Linear Regression

Introduction

This homework is one 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 does the last difference does the last problem highlight between ridge regression and standard least-squares regression?

Solution

We can expand the equation:

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

into

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

Where re-arranging we obtain:

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

(a)

(b)

(c)

(d)

(e)

Problem 2 (Priors and Regularization,7pts)

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

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

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

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

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.,

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

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 $\ln p(\mathbf{w} \mid \mathbf{y})$ 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

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.

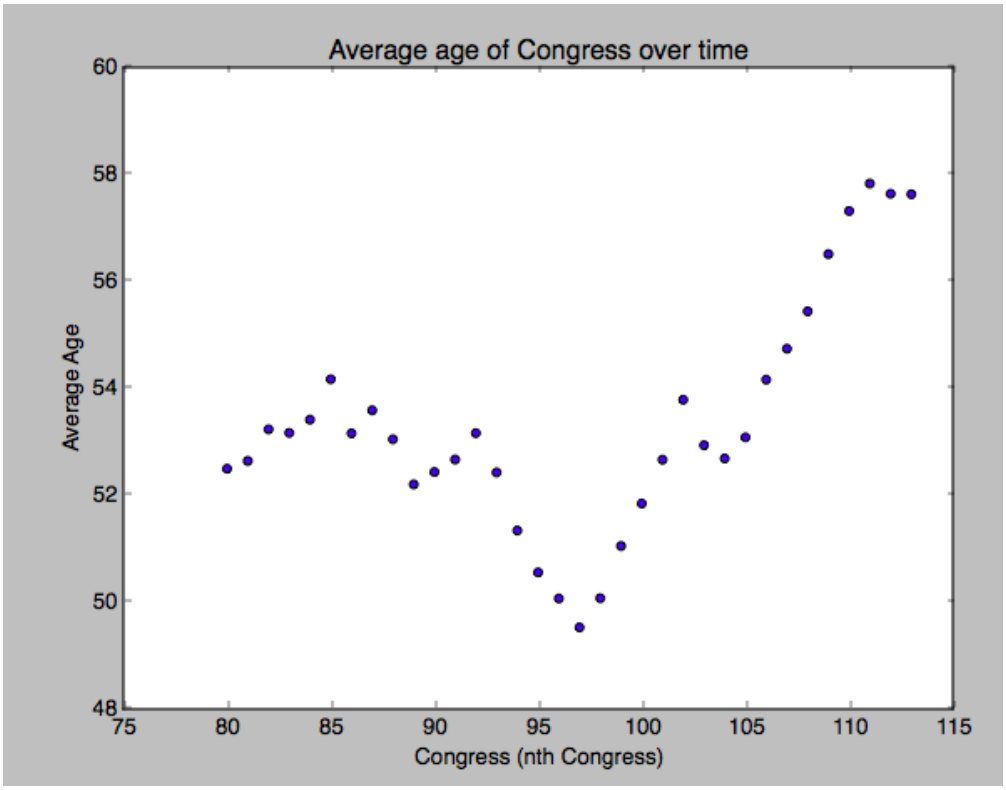

Figure 1: Average age of Congress. The horizontal axis is the Congress number, and the vertical axis is the average age of the congressmen.

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

(a)

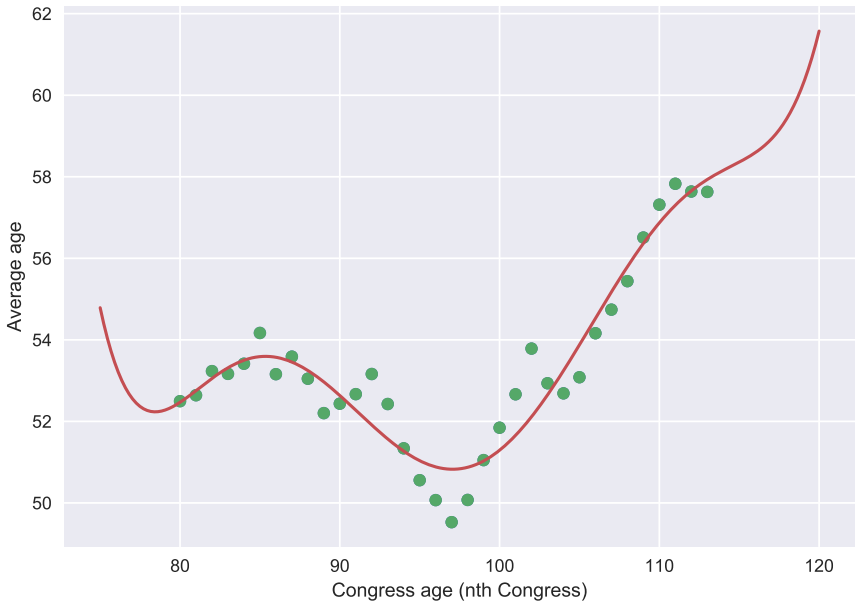

Figure 2

(b)

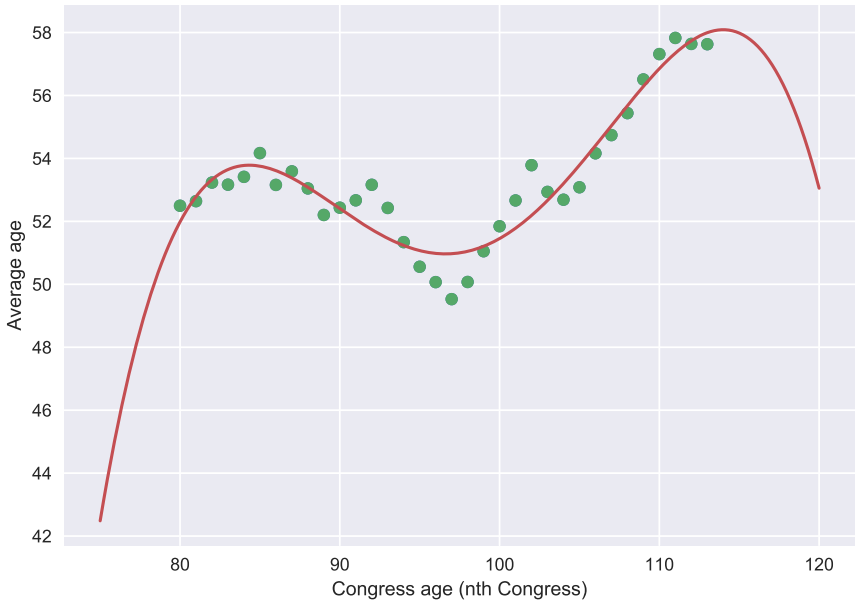

Figure 3

(c)

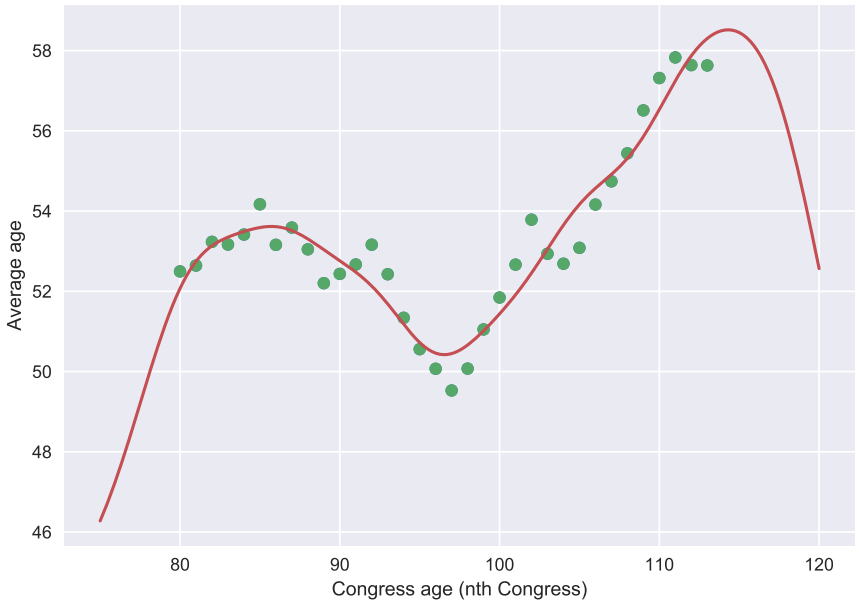

Figure 4

(d)

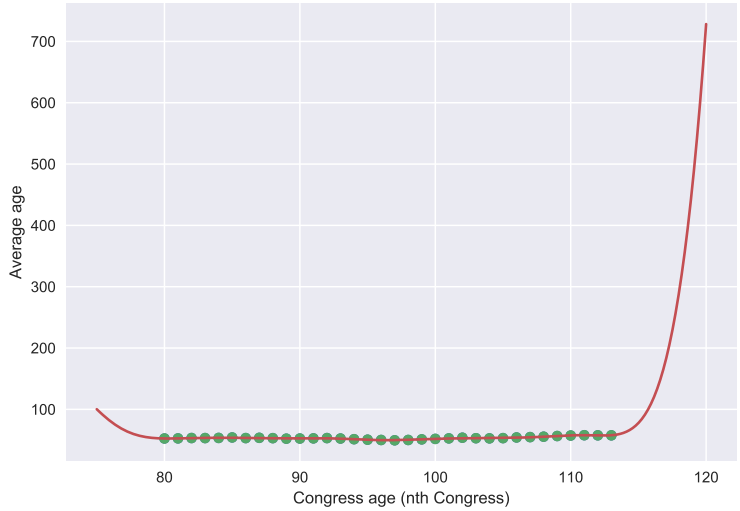

Figure 5

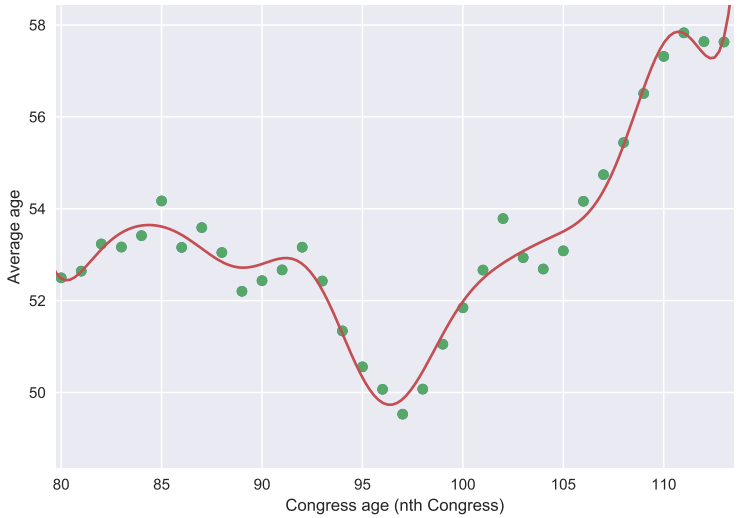

Figure 6

In fig. 5 we can see that the regression function does not do a very good job modeling the average age as a prediction for future congresses. However, from fig. 6, we can see that the model does a fairly good job capturing the age distribution for the data we have for the congress 80 to 115. I would trust this model to obtain data for the current congresses as a model, but would not trust it to make future predictions as to what the average age of congress will look like in the future.

(e)

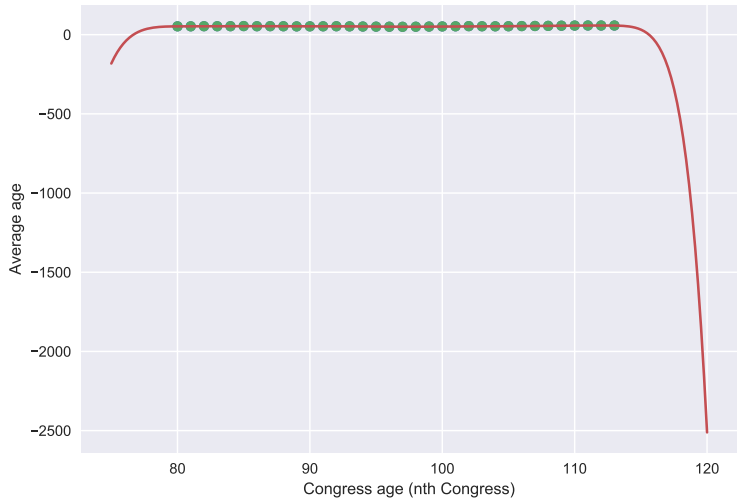

Figure 7

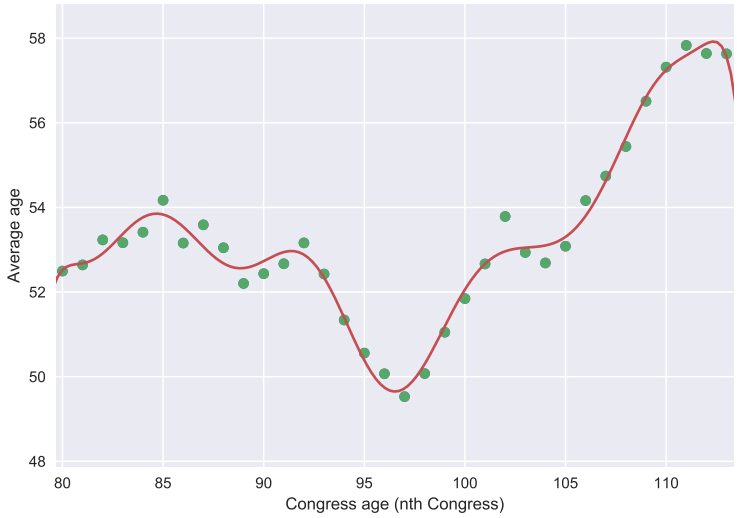

Figure 8

In this case we see a similar result as discussed in part (d), in that this model does a very good job capturing the features of the current data whereas it does not do a good job predicting the average age of the congresses in the future. Another feature to notice is that even though we added complexity to the basis functions, we did not gain that complexity back in form of accuracy. The plot does not show any significant benefit between what we were able to achieve in part (d) versus here. We can see that by just adding more terms to our basis functions, we cannot increase the accuracy of our model, but instead, if we want to make our models better we must choose different basis functions or a different model other than linear regression all together.

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

Answer: Approximately 4 hours.