Timothy Kang
tkang01@college.harvard.edu
CS181-S16

Assignment #1

Due: 5:00pm February 5, 2016

Collaborators:

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 $\boldsymbol{x} \in \mathbb{R}^n$ is centered, meaning $\forall \boldsymbol{x}, \sum_i x_i = 0$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ be the input matrix, the columns of which are the input vectors. 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

(a) We begin by noting that $(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})^T$ is a $1 \times n$ matrix, and that $(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})$ is an $n \times 1$ matrix. This indicates that the product of these two matrices is a 1×1 matrix. If we consider the example of the vector $[1 \quad 2 \quad 3]$ times its transpose, we would have the 1×1 matrix $[1^2 + 2^2 + 3^2]$. We observe a similar result when we multiply $(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})$ by its transpose: We would get a 1×1 matrix with entries

$$[(y_1 - \mathbf{x}_1\mathbf{w} - w_0)^2 + \cdots + (y_n - \mathbf{x}_n\mathbf{w} - w_0)^2]$$

This is equivalent to the expression

$$\sum_i (y_i - \mathbf{x}_1\mathbf{w} - w_0)^2$$

When we take the gradient w.r.t. w_0, we apply the chain rule and ignore the last term in $J(\mathbf{w}, w_0)$ since it does not contain w_0

$$\begin{aligned}\nabla_{w_0} J(\mathbf{w}, w_0) &= -2\sum_i (y_i - \mathbf{x}_1\mathbf{w} - w_0) \\ &= -2\left(\sum_i y_i - \sum_i \mathbf{x}_i\mathbf{w} - \sum_i w_0\right) \\ &= -2\left(\sum_i y_i - nw_0\right) \\ &= \boxed{2\left(nw_0 - \sum_i y_i\right)}\end{aligned}$$

We set the $\sum_i \mathbf{x}_i\mathbf{w}$ term equal to 0 because we use the fact that $\mathbf{x}$ is centered $\forall \mathbf{x}$, which by definition includes $\mathbf{x}$ multiplied by a factor - in this case, the coordinate of w_0.

(b) Before taking the gradient, we expand the polynomial and use properties of matrix transposes to simplify. Furthermore, we make use of the identity $\nabla_{\mathbf{z}}\mathbf{z}^T\mathbf{A}\mathbf{z} = (\mathbf{A} + \mathbf{A}^T)\mathbf{z}$

$$\begin{aligned}J(\mathbf{w}, w_0) &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{y}^T w_0\mathbf{1} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} \\ &\quad + \mathbf{w}^T\mathbf{X}^T w_0\mathbf{1} - \mathbf{1}^T w_0\mathbf{y} + \mathbf{1}^T w_0\mathbf{X}\mathbf{w} + \mathbf{1}^T w_0^2\mathbf{1} + \lambda\mathbf{w}^T\mathbf{w} \\ \nabla_{\mathbf{w}} J(\mathbf{w}, w_0) &= -2\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{I}\mathbf{w} \\ &= \boxed{-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{I}\mathbf{w}}\end{aligned}$$

We do not include the $\mathbf{1}^T w_0\mathbf{X}\mathbf{w}$ term and its transpose because they both result in 1×1 matrices with elements of the form

$$\mathbf{x}_1\mathbf{w}w_0 + \cdots + \mathbf{x}_n\mathbf{w}w_0$$

(transpose of 1×1 matrix is itself)
Using the same reasoning as part (a), we set this term equal to 0 because it results in a summation of the form

$$w_0\sum_i \mathbf{x}_i\mathbf{w}$$

(c)
- When we set the answer from (a) equal to zero (optimization entails setting the first derivative equal to 0), we get

$$\begin{aligned}2\left(nw_0 - \sum_i y_i\right) &= 0 \\ nw_0 &= \sum_i y_i \\ w_0 &= \frac{1}{n}\sum_i y_i\end{aligned}$$

- When we do the same for the answer for part (b), we get

$$\begin{aligned}-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{I}\mathbf{w} &= 0 \\ (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} &= \mathbf{X}^T\mathbf{y} \\ \mathbf{w} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

We prove that $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is invertible by demonstrating that it is positive definite $\forall v$

$$\begin{aligned}
v^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})v &= (v^T\mathbf{X}^T\mathbf{X} + \lambda v^T\mathbf{I})v \\
&= v^T\mathbf{X}^T\mathbf{X}v + \lambda v^T\mathbf{I}v \\
&= (\mathbf{X}v)^T(\mathbf{X}v) + \lambda(v^Tv) \\
&= \sum(\text{values})^2 + \lambda\sum(\text{values})^2 > 0
\end{aligned}$$

Therefore, $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is 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

We begin by observing that $p(\mathbf{w}|\alpha)$ is a multivariate normal distribution of dimensionality D with mean $\mathbf{0}$ and covariance matrix $\alpha^{-1}\mathbf{I}$. With this in mind, we plug in the appropriate values into the distribution for a multivariate normal (as given in the math review sheet).

$$\begin{aligned}
p(\mathbf{w}|\alpha) &= \frac{1}{\sqrt{2\pi^D \det(\alpha^{-1}\mathbf{I})}} e^{-\frac{1}{2}\mathbf{w}^T(\alpha^{-1}\mathbf{I})^{-1}\mathbf{w}} \\
\ln(p(\mathbf{w}|\alpha)) &= -\frac{1}{2}\mathbf{w}^T\mathbf{I}^{-1}\alpha\mathbf{w} + \ln\left(\frac{1}{\sqrt{2\pi^D \det(\alpha^{-1}\mathbf{I})}}\right) \\
&= -\frac{1}{2}\mathbf{w}^T\mathbf{w}\alpha + \text{constant} \\
&= -\alpha E_W(\mathbf{w}) + \text{constant}
\end{aligned}$$

We also observe that $p(\mathbf{t}|\mathbf{w})$ is a product of N univariate normal distributions with mean $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)$ and variance β^{-1}. We plug the values into the distribution for the univariate normal.

$$\begin{aligned}
p(\boldsymbol{t}|\boldsymbol{w}) &= \prod_{n=1}^{N} N(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\
&= \frac{1}{\sqrt{2\pi\beta^{-1}}}\prod_{n=1}^{N} e^{-\frac{1}{2\beta^{-1}}(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2} \\
\ln(p(\boldsymbol{t}|\boldsymbol{w})) &= -\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2 + \ln\left(\frac{1}{\sqrt{2\pi\beta^{-1}}}\right) \\
&= -\beta E_D(\mathbf{w}) + \text{constant}
\end{aligned}$$

Therefore, when we maximize the posterior, we are really maximizing

$$-(\alpha E_W(\mathbf{w}) + \beta E_D(\mathbf{w})) + \text{constant}$$

And if we remember that maximizing a function is the same as minimizing its negative, then it becomes clear that this is the same as minimizing

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

where $\lambda = \frac{\alpha}{\beta}$

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.

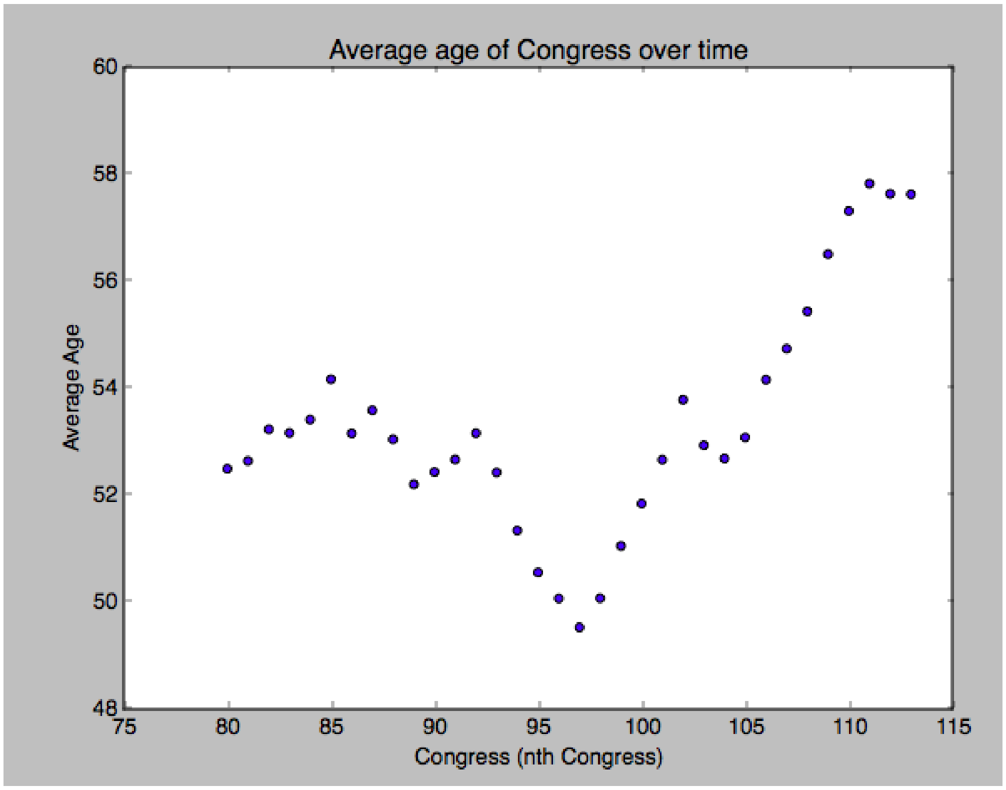


Figure 1: Average age of Congress. The horizontal axis is the Congress number, and the vertical axis is the average age of the congressmen.

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 $j = 1, \ldots, 4$

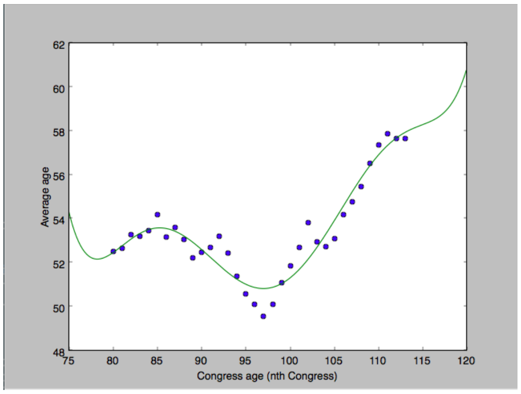
(d) $\phi_j(x) = \sin\{x/j\}$ for $j = 1, \ldots, 7$

(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.

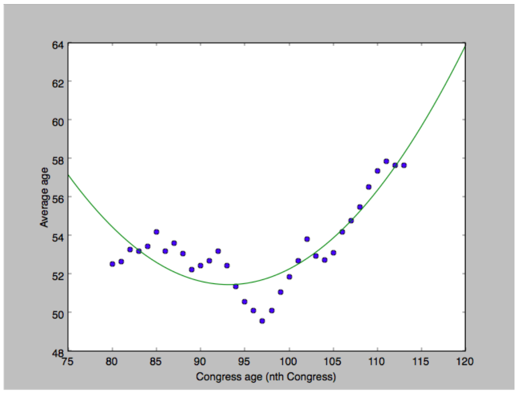
Solution

(a)



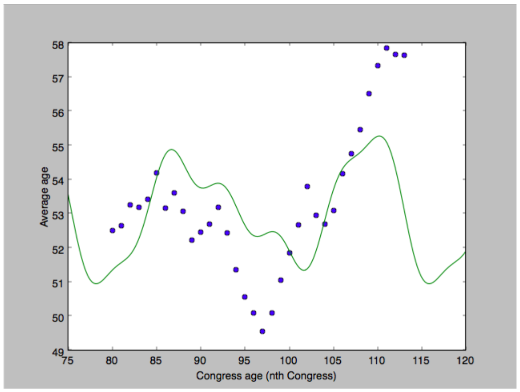
This regression is an overall good fit, as it tracks the trends in the data relatively accurately without major deviations, and it does not seem to be affected by potential outliers, such as those around the drop from 95-100 (on the x-axis).

(b)



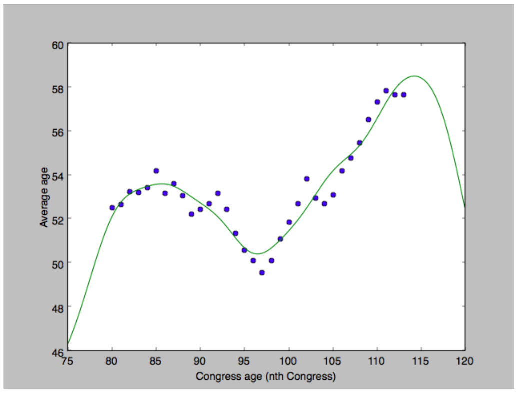
This regression under-fits the data, as it fails to account for most of the trends in the data, such as the peaks from 85-93 (on the x-axis) and the drop from 95-100 (on the x-axis).

(c)



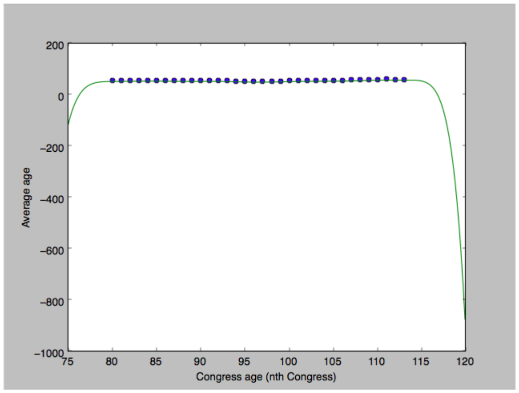
This one also under-fits the data because it does not follow the trends in the data. For example, it completely misses the peak from 100 - 104 (on the x-axis) and most of the drop from 95-100.

(d)



While this one seems to be a good fit (it is mostly similar to the regression from (a)), it still overfits the data by following the trends too closely. For example, the drop from 95 - 100 may constitute noise, but the regression follows the trend almost perfectly.

(e)



This regression heavily overfits the data. At this point, the regression is tracking not only the trends in the data but also all of the noise, as all of the points fit on the regression.

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

Answer: Roughly 3 hours. Most of it was remembering my linear algebra theory for problem 1.