Weiyi Chen
wec427@g.harvard.edu
CS181-S16

Assignment #1

Due: 5:00pm February 5, 2016

Collaborators: N.A.

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) The gradient of $J(w, w_0)$ with respect to w_0 is

$$\begin{aligned}
\nabla_{w_0} J(\boldsymbol{w}, w_0) &= 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1})^T \cdot (-\mathbf{1}) \\
&= -2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1})^T\mathbf{1} \\
&= -2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T\mathbf{1} + 2w_0 n \\
&= -2\boldsymbol{y}^T \cdot \mathbf{1} + \boldsymbol{w}^T\boldsymbol{X}^T\mathbf{1} + 2w_0 n \\
&= -2\sum_i y_i + \boldsymbol{w}^T\mathbf{0} + 2w_0 n \\
&= -2\sum_i y_i + 2w_0 n
\end{aligned}$$

(b) The gradient of $J(w, w_0)$ with respect to $\boldsymbol{w}$ is

$$\begin{aligned}\nabla_{\boldsymbol{w}} J(\boldsymbol{w}, w_0) &= 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1})^T \cdot (-\boldsymbol{X}) + 2\lambda \boldsymbol{w}^T \\ &= -2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - w_0\mathbf{1})^T \boldsymbol{X} + 2\lambda \boldsymbol{w}^T \\ &= 2\boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}) - 2\boldsymbol{y}^T\boldsymbol{X} + 2w_0\mathbf{1}^T\boldsymbol{X} \\ &= 2\boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}) - 2\boldsymbol{y}^T\boldsymbol{X} + 2w_0\mathbf{0} \\ &= 2\boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}) - 2\boldsymbol{y}^T\boldsymbol{X}\end{aligned}$$

(c) A global optimizer of $J(\boldsymbol{w}, w_0)$ is derived by letting the above gradients as zero, i.e.

$$\begin{aligned}\nabla_{w_0} J(\boldsymbol{w}, w_0) &= -2\sum_i y_i + 2w_0 n = 0 \Rightarrow w_0 = \frac{1}{n}\sum_i y_i \\ \nabla_{\boldsymbol{w}} J(\boldsymbol{w}, w_0) &= 2\boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}) - 2\boldsymbol{y}^T\boldsymbol{X} = \mathbf{0} \Rightarrow \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

In order to prove the inverse of $\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}$ exists, we can verify its determinant is non-zero, i.e.

$$det(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}) = |\boldsymbol{X}|^2 + \lambda > 0$$

given $|\boldsymbol{X}|^2 \geq 0$ and $\lambda > 0$, therefore $\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}$ 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

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

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

Write $\ln p(\boldsymbol{w} \mid \boldsymbol{t})$ as a function of $E_D(\boldsymbol{w})$ and $E_W(\boldsymbol{w})$, i.e.

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

where the multivariate normal distribution formula is

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

via substituting, we derive

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

Dropping the constant terms, we have

$$\begin{aligned} \ln p(\boldsymbol{w} \mid \boldsymbol{t}) - \boldsymbol{C} &= -\frac{1}{2}\alpha \boldsymbol{w}^T\boldsymbol{w} - \frac{1}{2}\beta \sum_{n=1}^{N}(t_n - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}_n))^2 \\ &= -\alpha E_W(\boldsymbol{w}) - \beta E_D(\boldsymbol{w}) \end{aligned}$$

Taking $\lambda = \alpha/\beta$,

$$\frac{1}{\beta}\left(\ln p(\boldsymbol{w} \mid \boldsymbol{t}) - \boldsymbol{C}\right) = -\lambda E_W(\boldsymbol{w}) - E_D(\boldsymbol{w})$$

Therefore, maximizing the posterior is equivalent to maximize $-\lambda E_W(\boldsymbol{w}) - E_D(\boldsymbol{w})$, or in other words, minimize the regularized error terms, i.e. $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$.

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.

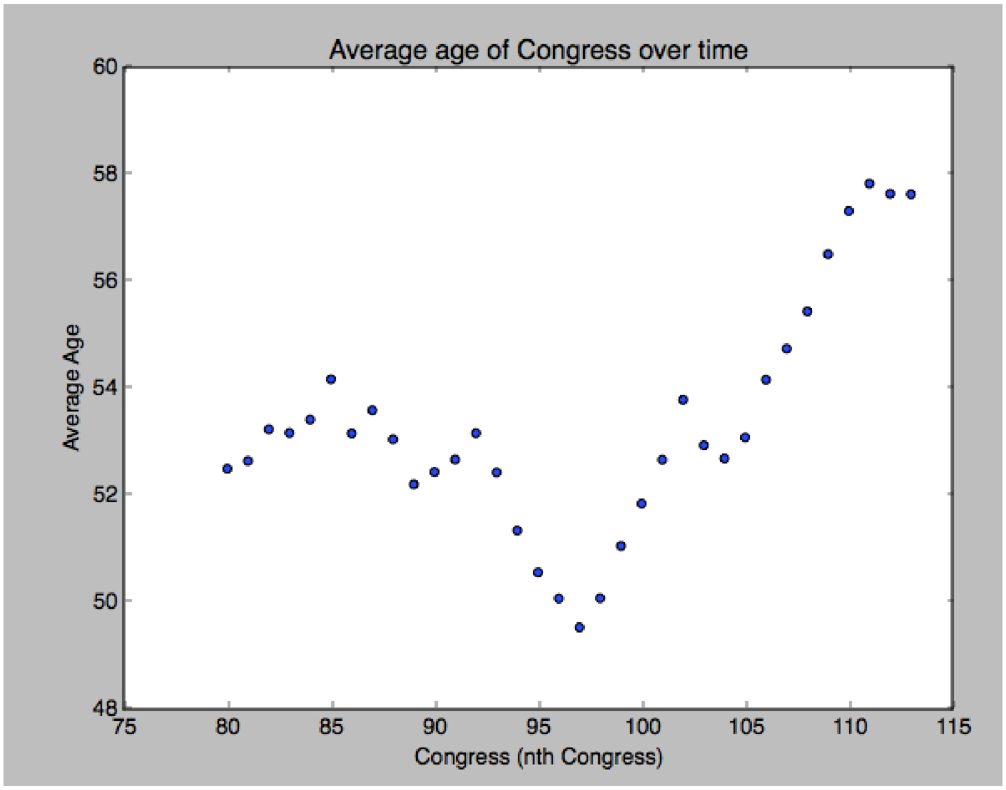

Figure 1: Average age of Congress. The horizontal axis is the Congress number, and the vertical axis is the average age of the congressmen.

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

Basis function regression is implemented in `linreg.py`. Below are the plots for each case -

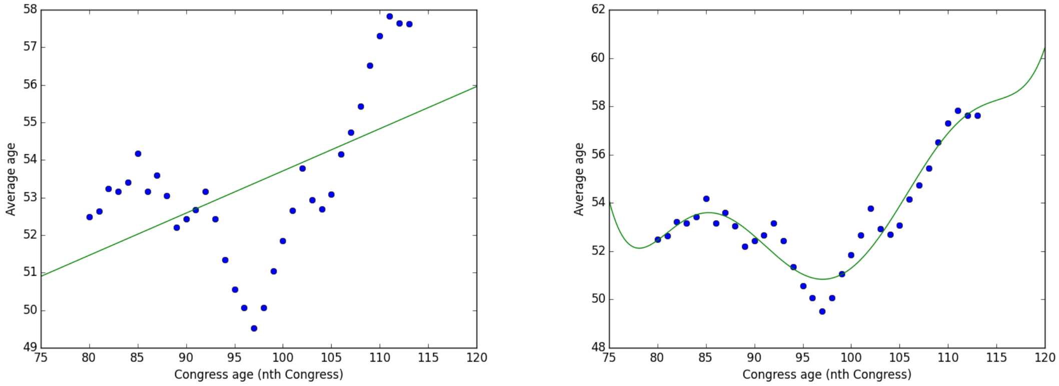

Figure 2: The data and regression line for the simple linear case and (a)

Following is my thinking of fitting well or not for each case -

- Fitting well: case (a), case (b), case (c), case (d)
- Overfitting: case (e)
- Underfitting: the simple linear case

The simple linear case does not fit well because it cannot capture the most important trends in the data, most of points are off from the line far away and the straight line cannot capture the curve indicated by the points either.

Case (e) is overfitting, though it captures every point exactly, it also fit to the error of each point, which is not expected. When the regression line goes out of the x range of the data points, the curve is going off far away to negative values very quickly, which is clearly unreasonable from our perspective.

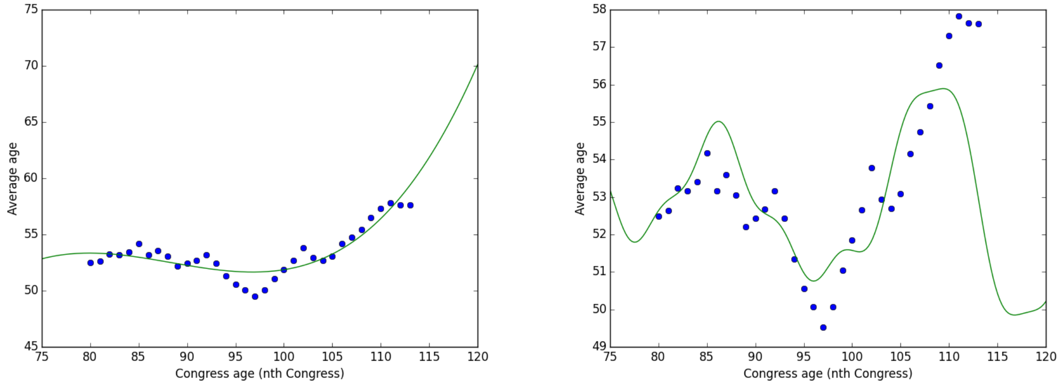

Figure 3: The data and regression line for cases (b) and (c)

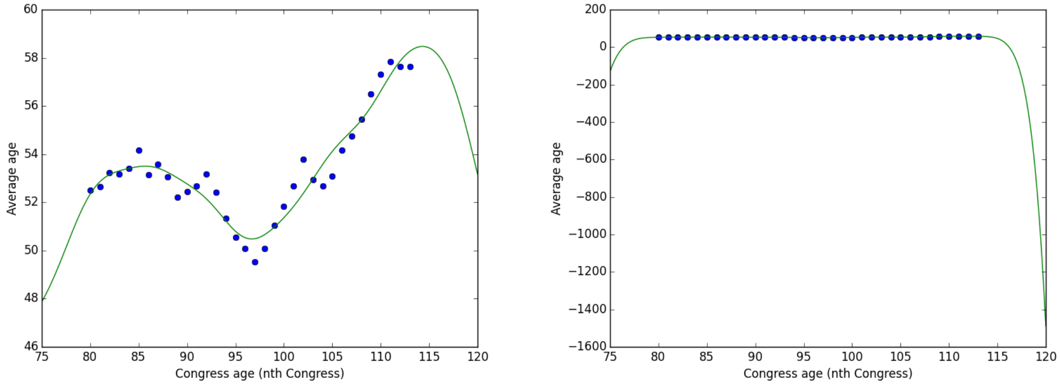

Figure 4: The data and regression line for cases (d) and (e)

Problem 4 (Calibration, 1pt)

Approximately how long did this homework take you to complete?

Answer: 3 hours. Roughly 1 hour for each question, including investigation time on unfamiliar knowledge.