Your Name
email@fas.harvard.edu
CS181-S16

Assignment #1
Due: 5:00pm February 5, 2016

Collaborators: John Doe, Fred Doe

Homework 1: Linear Regression, Solutions

Writeup due 23:59 on Friday 6 February 2015

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.

Grading Instructions: In the solutions, you will see several highlighted checkpoints. These each have a label that corresponds to an entry in the Google Form for this problem set. The highlighted statement should clearly indicate the criteria for being correct on that problem. If you satisfy the criteria for a problem being correct, mark "Yes" on the corresponding position on the Google Form. Otherwise, mark "No". Your homework scores will be verified by course staff at a later date.

Problem 1 (Centering and Ridge Regression, 7pts)

Consider a data set in which each data input vector $\boldsymbol{x} \in \mathbb{R}^m$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ be the input matrix, the rows of which are the input vectors, and the columns of which are centered at 0. 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 write the gradient 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

$$w_0 = \frac{1}{n}\sum_i y_i \tag{1}$$

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

For full credit, you will need to prove that each operation you perform is well-defined.

Solution 1

(a) Rewrite the original expression as such and compute the gradient of each term:

$$\underbrace{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda \boldsymbol{w}^T\boldsymbol{w}}_{\text{gradient is } 0} - \underbrace{w_0\mathbf{1}^T\boldsymbol{y} - w_0\boldsymbol{y}^T\mathbf{1}}_{\mathbf{1}^T\boldsymbol{y} = \boldsymbol{y}^T\mathbf{1} \text{ (scalar)}} + \underbrace{w_0\mathbf{1}^T(\boldsymbol{X}\boldsymbol{w}) + w_0(\boldsymbol{X}\boldsymbol{w})^T\mathbf{1}}_{\mathbf{1}^T\mathbf{X} = \mathbf{0} \ (\boldsymbol{X} \text{is centered})} + \underbrace{w_0^2\mathbf{1}^T\mathbf{1}}_{2w_0 n}$$

Check 1.1: Any combination of the following two expressions gives you full credit:

$$\nabla_{w_0} J(\boldsymbol{w}, w_0) = -2 \cdot \mathbf{1}^T\mathbf{y} + 2 \cdot w_0\mathbf{1}^T\mathbf{1} = -2\sum_i y_i + 2nw_0$$

(b) Rewrite the original expression as such and compute the gradient of each term:

$$\underbrace{\boldsymbol{y}^T\boldsymbol{y}}_{\text{gradient is 0}} \underbrace{-\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} - (\boldsymbol{X}\boldsymbol{w})^T\mathbf{y}}_{-2\boldsymbol{X}^T y} + \underbrace{\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}}_{2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}}$$
$$+ \underbrace{\lambda\boldsymbol{w}^T\boldsymbol{w}}_{2\lambda w} \underbrace{-\boldsymbol{y}^T w_0\mathbf{1} - w_0\mathbf{1}^T\mathbf{y} + w_0^2\mathbf{1}^\mathbf{T}\mathbf{1}}_{\text{gradient is 0}} \underbrace{-(\boldsymbol{X}\boldsymbol{w})^T w_0\mathbf{1} - w_0\mathbf{1}^T\boldsymbol{X}\boldsymbol{w}}_{\mathbf{1}^T\mathbf{X}=0}$$

Hint: To compute gradients in matrix form, the best thing to do is usually to replace the variable x you are differentiating with respect to by $x + \epsilon$, and identify the gradient as the vector that ϵ is taken a dot product of:

$$f(x_0 + \epsilon) = f(x_0) + \epsilon^T \underbrace{\nabla_x f(x_0)}_{\text{Identify the gradient here}} + (\text{higher order terms in } \epsilon)$$

Check 1.2: The following expression gives you full credit (up to reordering of the terms):

$$\nabla_{\boldsymbol{w}} J(\boldsymbol{w}, w_0) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + 2\lambda\boldsymbol{w}$$

(c) Function J is convex in $(\boldsymbol{w}, w_0)$ and therefore any local minima is also a global minima. Furthermore, by convexity, any point $(\boldsymbol{w}^*, w_0^*)$ where the gradient is 0 is a local minima. Solving for

$$\nabla_{w_0} J(\boldsymbol{w}, w_0) = 0 \Leftrightarrow -2\sum_i y_i + 2nw_0 = 0 \Leftrightarrow w_0 = \frac{1}{n}\sum_i y_i$$
$$\nabla_{w} J(\boldsymbol{w}, w_0) = 0 \Leftrightarrow -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + 2\lambda\boldsymbol{w} = 0 \Leftrightarrow \boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)\boldsymbol{w} = 0 \Leftrightarrow \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T y$$

yields the above solution.

It is necessary to justify that $X^T X + \lambda I$ is invertible when $\lambda > 0$. There are two main ways to justify that this matrix is invertible:

1. Note that $\forall x \in \mathbb{R}^n, \boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda I)\boldsymbol{x} = \|\boldsymbol{X}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2$, which is positive and equal to 0 iff $\boldsymbol{x} = 0$.
2. State that $\boldsymbol{X}^T\boldsymbol{X}$ has non-negative eigenvalues (or is semi-positive definite), commutes with the identity matrix, that they are therefore simultaneously diagonizable, and that their sum has strictly positive eigenvalues. Note that $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible in all generality and making that assumption does not grant you points for Check 1.5.

To get full points for this question, you needed to:

- **Check 1.3**: state that the optimum is found by solving for each gradient in 1.a and 1.b equal to 0
- **Check 1.4**: solve the 2 matrix equations
- **Check 1.5**: correctly justify that $(X^T X + \lambda I)$ is invertible when $\lambda > 0$.

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: $p(\boldsymbol{w} \mid \alpha)$ is a multivariate normal distribution. Suppose the dimension of $\boldsymbol{w}$ is $D \times 1$. Plug the mean $\mathbf{0}$ and covariance matrix $\alpha^{-1}\boldsymbol{I}$ into the PDF of multivariate normal distribution:

$$p(\boldsymbol{w} \mid \alpha) = \frac{1}{\sqrt{(2\pi)^D det(\alpha^{-1}\boldsymbol{I})}} exp(-\frac{1}{2}\boldsymbol{w}^T(\alpha^{-1}\boldsymbol{I})^{-1}\boldsymbol{w})$$
$$\ln p(\boldsymbol{w} \mid \alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}\alpha + constant = -\alpha E_W(\boldsymbol{w}) + constant$$

Check 2.1: You correctly took the log of $p(w|\alpha)$. It is acceptable to write out the constants, or just write out "+ constant" as done in the solution.

Similarly,

$$p(\boldsymbol{t} \mid \boldsymbol{w}) = \frac{1}{\sqrt{2\beta^{-1}\pi}} \prod_{n=1}^{N} exp(-\frac{(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2}{2\beta^{-1}})$$
$$\ln p(\boldsymbol{t} \mid \boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2\beta + constant = -\beta E_D(\boldsymbol{w}) + constant$$

Check 2.2: You correctly take the log of $p(t|w)$. It is acceptable to write out the constants, or just write out "+ constant" as done in the solution.

Therefore, maximizing $\ln p(\boldsymbol{w}|\alpha) + \ln p(\boldsymbol{t} \mid \boldsymbol{w})$ is equivalent to maximizing $-\beta E_D(\boldsymbol{w}) - \alpha E_W(\boldsymbol{w})$. Hence it is equal to minimizing $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$, where $\lambda = \alpha/\beta$. (Note that $\beta > 0$ because it is a variance.).

Check 2.3: You correctly show how the sum of those two is equal to minimizing $E_D(w) + \lambda E_W(w)$.

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. Download the file `congress-ages.csv` from the course website. It has two columns. The first one is an integer that indicates the Congress number. We are currently in the 114th US Congress. 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.

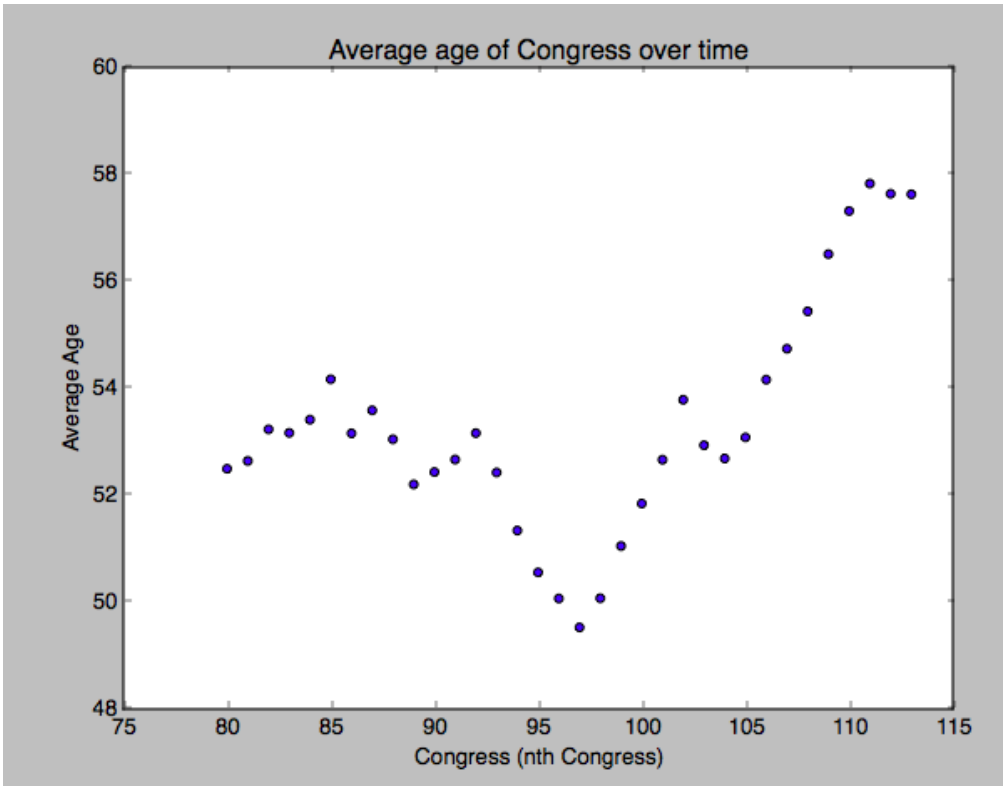


Figure 1: Average age of Congress. The horizontal axis is the Congress number, and the vertical axis is the average age of the congressmen.

Problem 3 (Modeling Changes in Congress, 10pts)

Implement basis function regression with ordinary least squares. Some sample Python code is provided in `linreg.py`. 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 $j = 1, \ldots, 3$

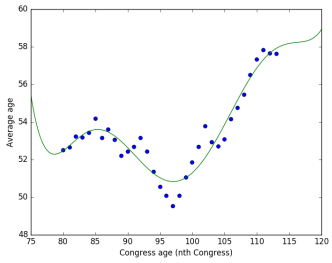
(c) $\phi_j(x) = \sin\{x/j\}$ for $j = 1, \ldots, 4$

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

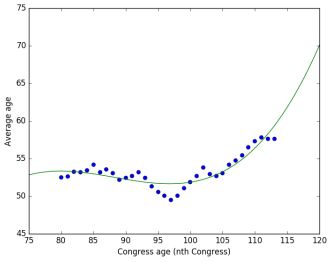
(e) $\phi_j(x) = \sin\{x/j\}$ for $j = 1, \ldots, 20$

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.

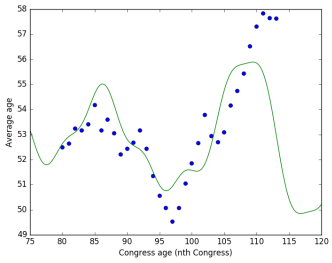
Solution: See the posted solution code.



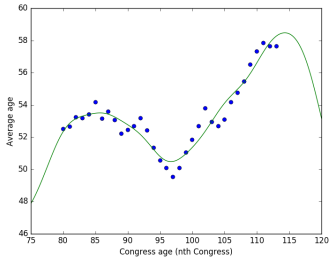
(a)



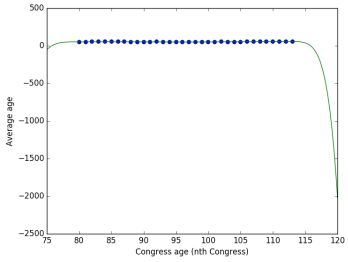
(b)



(c)



(d)



(e)

- **Check 3.1**: Your graph should match the graph for (a) above.
- **Check 3.2**: Your graph should match the graph for (b) above.
- **Check 3.3**: Your graph should match the graph for (c) above.
- **Check 3.4**: Your graph should match the graph for (d) above.
- **Check 3.5**: Your graph should match the graph for (e) above.

(a) and (d) fit relatively well, though (a) may siightly underfit (both answers are acceptable). (b) and (c) exhibit underfitting: (b) seems to fail to capture the fall in age towards the middle of the plot, and (c) does not fit the plot well towards the right side. This may be because the basis functions in these two are not sufficiently high-dimensional to capture the interesting parts of the dataset. (e) is overfitting: it fits the given data well, but also has captured the noise. The steep rising / falling tails are not likely patterns of the original data. This is because we fit the data to a high-degree basis function (20th degree).

- **Check 3.6**: Your explanation for (a) should be that it was either a good fit, or that it slightly underfits.
- **Check 3.7**: Your explanation for (b) should be that it underfits, for example because it fails to capture the fall in age towards the middle of the plot, perhaps because there are not enough dimensions.
- **Check 3.8**: Your explanation for (c) should be that it underfits, as it does not fit the plot well towards the right side. The basis functions are not sufficiently high-dimensional.
- **Check 3.9**: Your explanation for (d) should be that it was a good fit.
- **Check 3.10**: Your explanation for (e) should be that it overfits, as explained by the extreme values on either side. This is because the data is very high dimensional, and is modeling the noise.

Calibration [1pt]

Approximately how long did this homework take you to complete?

Changelog

- **v1.0** – 28 January 2015 at 13:00
- **v1.1** – 28 January 2015 at 22:30. Removed extra λ in $E_W(\boldsymbol{w})$, minor formatting edits
- **v1.2** – 21 January 2016 at 15:20. Updated to new problems and added grading guidelines throughout.