Your Name
email@fas.harvard.edu
CS181-S17

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

Collaborators: John Doe, Fred Doe

Homework 1 Solutions

Grading Instructions

In the solutions, you will see several highlighted checkpoints. These each have a label that corresponds to an entry in the Canvas quiz 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 Canvas quiz. 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 $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ 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

$$w_0 = \frac{1}{n}\sum_{i=1}^n y_i \qquad (1)$$

$$\mathbf{w} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \qquad (2)$$

(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 between standard least-squares regression versus ridge regression?

Solution

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

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

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

$$\frac{d\mathcal{L}(\mathbf{w}, w_0)}{dw_0} = -2\cdot\mathbf{1}^\top\mathbf{y} + 2\cdot w_0\mathbf{1}^\top\mathbf{1} = -2\sum_i y_i + 2nw_0$$

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

$$\underbrace{\mathbf{y}^\top\mathbf{y}}_{\text{gradient is 0}} - \underbrace{\mathbf{y}^\top\mathbf{X}\mathbf{w} - (\mathbf{X}\mathbf{w})^\top\mathbf{y}}_{-2\mathbf{X}^\top y} + \underbrace{\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}}_{2\mathbf{X}^\top\mathbf{X}\mathbf{w}}$$

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

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

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

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

$$\frac{\delta \mathcal{L}(\mathbf{w}, w_0)}{\delta w_0} = 0 \Leftrightarrow -2\sum_i y_i + 2nw_0 = 0 \Leftrightarrow w_0 = \frac{1}{n}\sum_i y_i$$

$$\frac{\delta \mathcal{L}(\mathbf{w}, w_0)}{\delta \mathbf{w}} = 0 \Leftrightarrow -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\mathbf{w} + 2\lambda \mathbf{w} = 0 \Leftrightarrow \mathbf{X}^\top \mathbf{y} - (\mathbf{X}^\top \mathbf{X} + \lambda I)\mathbf{w} = 0 \Leftrightarrow \mathbf{w} = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top y$$

yields the above solution.

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

- **Check 1.3**: state that the optimum found by solving for each gradient in 1.a and 1.b equal to 0
- **Check 1.4**: solve the 2 matrix equations

(d) A symmetric $m \times m$ matrix $\mathbf{A}$ is positive semi-definite if for all non-zero $\mathbf{z} \in \mathbb{R}^m$ we have $\mathbf{z}^\top \mathbf{A}\mathbf{z} \geq 0$. A symmetric matrix $\mathbf{A}$ is positive definite (and thus full rank, and thus invertible) if for all non-zero $\mathbf{z} \in \mathbb{R}^m$, we have $\mathbf{z}^\top \mathbf{A}\mathbf{z} > 0$. Now, we have $\mathbf{z}^\top(\mathbf{A} + \lambda \mathbf{I})\mathbf{z} = \mathbf{z}^\top \mathbf{A}\mathbf{z} + \lambda \mathbf{z}^\top \mathbf{z} = \mathbf{z}^\top \mathbf{A}\mathbf{z} + \lambda ||\mathbf{z}||^2 > 0$, since $\lambda > 0$ and $\mathbf{z}$ is non-zero.

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

- **Check 1.5**: correctly justify that $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})$ is invertible when $\lambda > 0$.

(e) Adding the $\lambda \mathbf{I}$ regularization term guarantees positive-definiteness whereas $\mathbf{X}^\top \mathbf{X}$ by itself does not. This ensures that an analytical term is possible, as long as $\lambda > 0$.

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

- **Check 1.6**: correctly justify that $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})$ guarantees invertibility whereas $(\mathbf{X}^\top \mathbf{X})$ without the regularization term does not.

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,

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

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.,

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

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 $\ln p(\mathbf{w} \,|\, \mathbf{y}, \mathbf{X})$ 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

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

$$p(\mathbf{w}) = \frac{1}{(|2\pi\Sigma|)^{1/2}}\exp(-\frac{1}{2}\mathbf{w}^\top(\alpha^{-1}\mathbf{I})^{-1}\mathbf{w})$$
$$\ln p(\mathbf{w}) = -\frac{1}{2}\mathbf{w}^\top\mathbf{w}\alpha + \mathit{constant} = -\alpha R(\mathbf{w}) + \mathit{constant}$$

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

Similarly,

$$p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}) = \frac{1}{\sqrt{2\beta^{-1}\pi}}\prod_{i=1}^{n}\exp(-\frac{(y_i - \mathbf{w}^\top\mathbf{x}_i)^2}{2\beta^{-1}})$$
$$\ln p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top\mathbf{x}_i)^2\beta + \mathit{constant} = -\beta L(\mathbf{w}) + \mathit{constant}$$

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

Therefore, maximizing $\ln p(\mathbf{w}) + \ln p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})$ is equivalent to maximizing $-\beta L(\mathbf{w}) - \alpha R(\mathbf{w})$. Hence it is equal to minimizing $L(\mathbf{w}) + \lambda R(\mathbf{w})$, where $\lambda = \alpha/\beta$. (Note that $\beta > 0$ because it is a variance.).

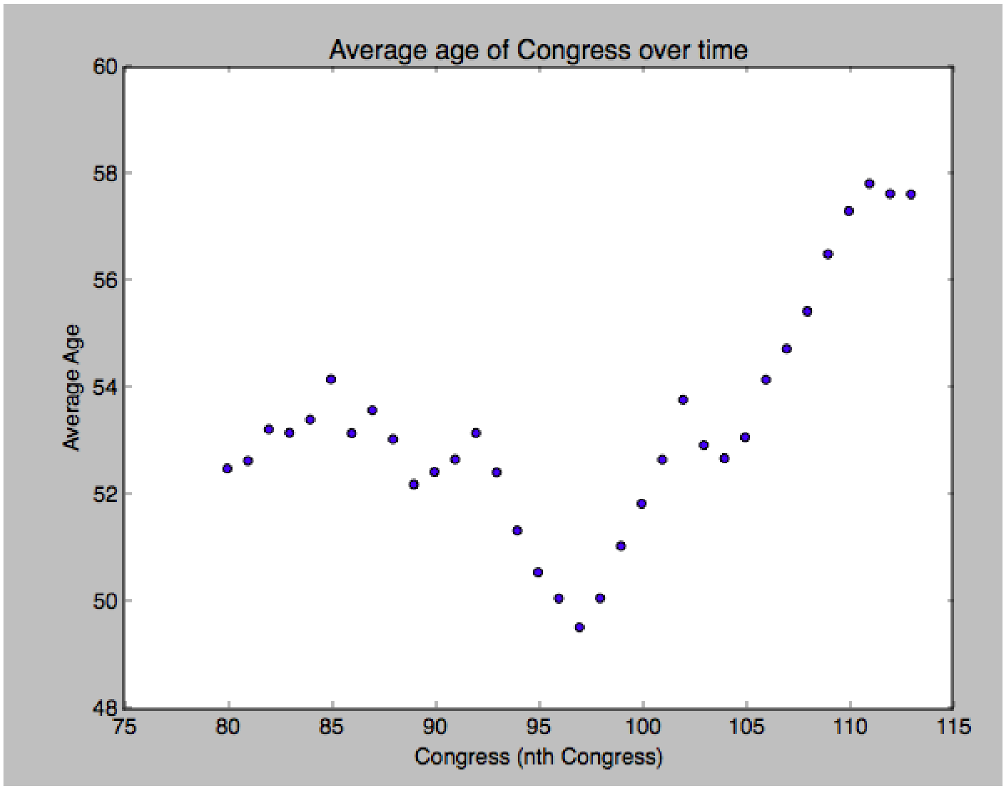
Check 2.3: You correctly show how the sum of those two is equal to minimizing $L(\mathbf{w}) + \lambda R(\mathbf{w})$.

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 ??.



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 $j = 1, \ldots, 4$

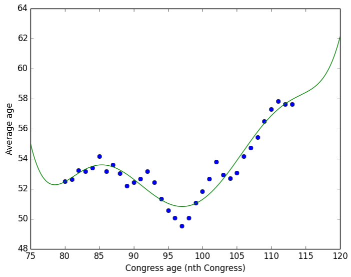
(c) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 6$

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

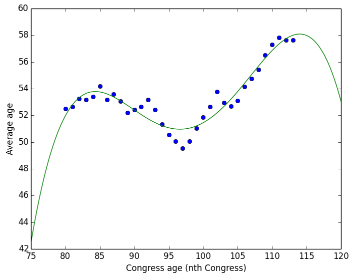
(e) $\phi_j(x) = \sin(x/j)$ for $j = 1, \ldots, 22$

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.

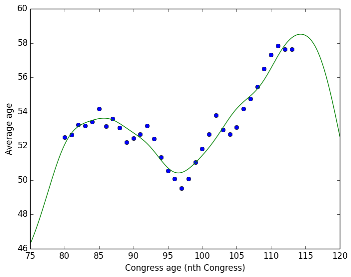
Solution



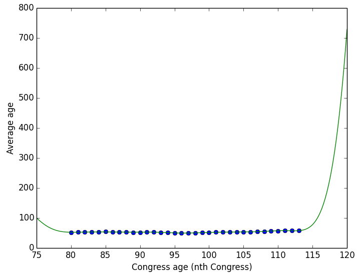
(a)



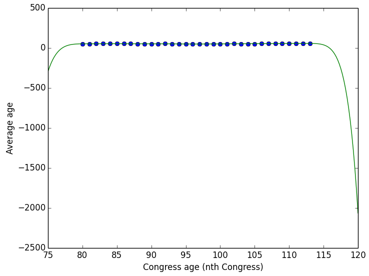
(b)



(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) fits relatively well/slightly underfits (both answers are acceptable). It has a loss of 6.56, which is a significant improvement over the linear basis loss of 55.30. However, it does not completely capture the trough in the middle of the plot, suggesting a slight underfit.

(b) fits relatively well/slightly underfits (both answers are acceptable). It has a loss of 7.02 and also does not completely the trough.

(c) fits relatively well/slightly overfits (both answers are acceptable). It has a loss of 5.71, which is lower than that of (a) and (b) which suggests that it comes close to an overfit. We can indeed see graphically that the regression line loses some of its smoothness, suggesting it is starting to slightly overfit to the noise of the data.

(d) and (e) are overfitting: it fits the given data well, but also has captured the noise. They have losses of 1.93 and 1.45 respectively, which are much lower than that of (a), (b) and (c), suggesting an overfit. The steep rising and falling tails are not likely patterns of the original data. This is because we fit the data to a high-degree basis function (10th and 22th degree).

- **Check 3.6**: Your explanation for (a) should be that it is either a good fit, or that is slightly underfits, for example because it fails to fully capture the trough towards the middle of the plot, perhaps because there are not enough dimensions. Must include a loss of 6.56.

- **Check 3.7**: Your explanation for (b) should be that it is either a good fit, or that is slightly underfits. Must include a loss of 7.02.

- **Check 3.8**: Your explanation for (c) should be that it is a good fit, since it seems to match the data well without overly hugging the points. You could also point out that there is strange behavior towards the extremes, which suggests a slight overfit. Must include a loss of 5.71.

- **Check 3.9**: Your explanation for (d) should be that it was an overfit. Must include a loss of 1.93.

- **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. Must include a loss of 1.45.

Problem 4 (Calibration, 1pt)

Approximately how long did this homework take you to complete?

Answer: