Theory of Neural Networks Mini Project 1

Michael Person

March 26, 2017

Contents

List of Figures

1 Task 1

1.1 Part A

From the lecture notes, we have the Covariance matrix of the parameters and can use that to get the precision matrix.

$$\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N) = \left(s^2\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1} \tag{1}$$

$$\mathbf{C}^{-1}_{\mathbf{w}|\mathbf{t}}(N) = s^2\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} \tag{2}$$

Next using equation 1.25 from the lecture notes:

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i^T\mathbf{x}_i \tag{3}$$

We can rewrite the precision matrix as the following:

$$\mathbf{C}^{-1}_{\mathbf{w}|\mathbf{t}}(N) = s^2\mathbf{I} + \frac{1}{\sigma^2}\sum_{i=1}^{N} \mathbf{x}_i^T\mathbf{x}_i \tag{4}$$

Since the $s^2\mathbf{I}$ term does not change with the number of inputs because it is based upon the prior distribution of our parameters we are able to replace N with $N+1$ and make the following equation:

$$\mathbf{C}^{-1}_{\mathbf{w}|\mathbf{t}}(N+1) = s^2\mathbf{I} + \frac{1}{\sigma^2}\sum_{i=1}^{N+1} \mathbf{x}_i^T\mathbf{x}_i \tag{5}$$

$$= s^2\mathbf{I} + \frac{1}{\sigma^2}\sum_{i=1}^{N} \mathbf{x}_i^T\mathbf{x}_i + \frac{1}{\sigma^2}\mathbf{x}_{N+1}^T\mathbf{x}_{N+1} \tag{6}$$

Therefor we have the following equation:

$$\mathbf{C}^{-1}_{\mathbf{w}|\mathbf{t}}(N+1) = \mathbf{C}^{-1}_{\mathbf{w}|\mathbf{t}}(N) + \frac{1}{\sigma^2}\mathbf{x}_{N+1}^T\mathbf{x}_{N+1} \tag{7}$$

1.2 Part B

From Part A, we derived the precision matrix 7 and can be used to solve for the covariance:

$$\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1) = \left(\mathbf{C}^{-1}_{\mathbf{w}|\mathbf{t}}(N) + \frac{1}{\sigma^2}\mathbf{x}_{N+1}^T\mathbf{x}_{N+1}\right)^{-1} \tag{8}$$

Then using the Sherman-Morrison Formula, which is a special case of the Woodbury Matrix Identity, from [1] we can rewrite the covariance matrix:

$$\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1) = \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N) - \frac{\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\frac{1}{\sigma^2}\mathbf{x}_{N+1}^T\mathbf{x}_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)}{1 + \frac{1}{\sigma^2}\mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}} \tag{9}$$

$$= \mathbf{I}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N) - \frac{\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}^T\mathbf{x}_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)}{\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}} \tag{10}$$

$$= \left(\mathbf{I} - \frac{\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}^T\mathbf{x}_{N+1}}{\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}\right)\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N) \tag{11}$$

Therefore we have the following solution for updating our covariance:

$$\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1) = \mathbf{G}(N+1)\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N) \tag{12}$$

$$\mathbf{G}(N+1) = \mathbf{I} - \frac{\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}^T\mathbf{x}_{N+1}}{\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}} \tag{13}$$

1.3 Part C

Now we need to solve for the updated posterior of the parameters with the new data point that we have received. From the notes we know that:

$$\widehat{\mathbf{w}}_{MAP}(N) = \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\left[\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \mathbf{S}^{-1}\mathbf{w}_0\right] \tag{14}$$

$$= \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\left[\frac{1}{\sigma^2}\sum_{i=1}^{N} t_i\mathbf{x}_i + \mathbf{S}^{-1}\mathbf{w}_0\right] \tag{15}$$

And for the next data point becomes:

$$\widehat{\mathbf{w}}_{MAP}(N+1) = \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\left[\frac{1}{\sigma^2}\sum_{i=1}^{N+1} t_i\mathbf{x}_i + \mathbf{S}^{-1}\mathbf{w}_0\right] \tag{16}$$

$$= \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\left[\frac{1}{\sigma^2}\sum_{i=1}^{N} t_i\mathbf{x}_i + \mathbf{S}^{-1}\mathbf{w}_0 + \frac{1}{\sigma^2}t_{N+1}\mathbf{x}_{N+1}\right] \tag{17}$$

$$= \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\left[\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \mathbf{S}^{-1}\mathbf{w}_0\right] + \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\frac{1}{\sigma^2}t_{N+1}\mathbf{x}_{N+1} \tag{18}$$

Then using equation 12, we can rewrite:

$$\widehat{\mathbf{w}}_{MAP}(N+1) = \mathbf{G}(N+1)\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\left[\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \mathbf{S}^{-1}\mathbf{w}_0\right] + \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\frac{1}{\sigma^2}t_{N+1}\mathbf{x}_{N+1} \tag{19}$$

$$= \mathbf{G}(N+1)\widehat{\mathbf{w}}_{MAP}(N) + \mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\frac{1}{\sigma^2}t_{N+1}\mathbf{x}_{N+1} \tag{20}$$

Now we look at the second term and massage it:

$$\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\frac{1}{\sigma^2}t_{N+1}\mathbf{x}_{N+1} = \frac{1}{\sigma^2}t_{N+1}\left(\mathbf{I} - \frac{\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}^T\mathbf{x}_{N+1}}{\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}\right)\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1} \tag{21}$$

$$= \frac{t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}{\sigma^2} - \frac{t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}\mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}{\sigma^2\left(\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}\right)} \tag{22}$$

Looking at this first term:

$$\frac{t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}{\sigma^2} = \frac{t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}\left(\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}\right)}{\sigma^2\left(\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}\right)} \tag{23}$$

$$= \frac{\sigma^2 t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1} + t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}\mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}{\sigma^2\left(\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}\right)} \tag{24}$$

Then plugging back into equation 22 and canceling the same terms, we get the following equation:

$$\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N+1)\frac{1}{\sigma^2}t_{N+1}\mathbf{x}_{N+1} = \frac{\sigma^2 t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}{\sigma^2\left(\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}\mathbf{x}_{N+1}\right)} \tag{25}$$

$$= \frac{t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}{\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}\mathbf{x}_{N+1}} \tag{26}$$

Then plugging this term back into our initial equation 20:

$$\widehat{\mathbf{w}}_{MAP}(N+1) = \mathbf{G}(N+1)\widehat{\mathbf{w}}_{MAP}(N) + \frac{t_{N+1}\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}}{\sigma^2 + \mathbf{x}_{N+1}^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}_{N+1}} \tag{27}$$

1.4 Part D

We wish to initialize the MAP parameter mean and covariance before seeing any data in such a way that we will get a correct algorithm. We start from the definition of the MAP parameter and find it after a single point has been found:

$$\widehat{\mathbf{w}}(1) = \left(\mathbf{x}_1\mathbf{x}_1^T + \mu s^2\mathbf{I}\right)^{-1}\mathbf{x}_1 t_1 \tag{28}$$

By the Sherman-Morrison formula we can rewrite that as the following:

$$\widehat{\mathbf{w}}(1) = (\mu s^2 \mathbf{I})^{-1} - \frac{(\mu s^2 \mathbf{I})^{-1} \mathbf{x}_1 \mathbf{x}_1^T (\mu s^2 \mathbf{I})^{-1}}{1 + \mathbf{x}_1^T (\mu s^2 \mathbf{I})^{-1} \mathbf{x}_1} \tag{29}$$

And we will define $A = \mu s^2$

$$\begin{aligned}
\widehat{\mathbf{w}}(1) &= A\mathbf{I}\mathbf{x}_1 t_1 - \frac{A\mathbf{I}\mathbf{x}_1 \mathbf{x}_1^T A\mathbf{I}\mathbf{x}_1 t_1}{1 + A\mathbf{x}_1^T \mathbf{I}\mathbf{x}_1} && (30)\\
&= A\mathbf{x}_1 t_1 - \frac{A^2 \mathbf{x}_1 \mathbf{x}_1^T \mathbf{x}_1 t_1}{1 + A\mathbf{x}_1^T \mathbf{x}_1} && (31)\\
&= \frac{A\mathbf{x}_1 t_1 + A^2 \mathbf{x}_1 \mathbf{x}_1^T \mathbf{x}_1 t_1 - A^2 \mathbf{x}_1 \mathbf{x}_1^T \mathbf{x}_1 t_1}{1 + A\mathbf{x}_1^T \mathbf{x}_1} && (32)\\
\widehat{\mathbf{w}}(1) &= \frac{A\mathbf{x}_1 t_1}{1 + A\mathbf{x}_1^T \mathbf{x}_1} && (33)
\end{aligned}$$

Next we use our update equation found in Equation 27 to find what the update from the initialization to our MAP parameter after the first datapoint has been inputted:

$$\begin{aligned}
\widehat{\mathbf{w}}(1) &= \mathbf{G}(1)\widehat{\mathbf{w}}(0) + \frac{t_1 \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1}{\sigma^2 + \mathbf{x}_1^T \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1} && (34)\\
&= \left(\mathbf{I} - \frac{\mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1 \mathbf{x}_1^T}{\sigma^2 + \mathbf{x}_1^T \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1}\right)\widehat{\mathbf{w}}(0) + \frac{t_1 \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1}{\sigma^2 + \mathbf{x}_1^T \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1} && (35)\\
&= \frac{\widehat{\mathbf{w}}(0)(\sigma^2 + \mathbf{x}_1^T \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1) - \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1 \mathbf{x}_1^T \widehat{\mathbf{w}}(0) + t_1 \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1}{\sigma^2 + \mathbf{x}_1^T \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1} && (36)\\
&= \frac{\widehat{\mathbf{w}}(0)\sigma^2 \mathbf{x}_1^T \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1 - \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1 \mathbf{x}_1^T \widehat{\mathbf{w}}(0) + t_1 \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1}{\sigma^2 + \mathbf{x}_1^T \mathbf{C}_{\mathbf{w}|\mathbf{t}}(0)\mathbf{x}_1} && (37)
\end{aligned}$$

Matching the denominators of the above equation and Equation 33 we find that the initialization for the covariance of the parameters is the Identity $\mathbf{C}_{\mathbf{w}|\mathbf{t}}(0) = \mathbf{I}$. Plugging this into Equation 37 and simplifying:

$$\begin{aligned}
\widehat{\mathbf{w}}(1) &= \frac{\sigma^2}{\sigma^2} \frac{\widehat{\mathbf{w}}(0) + \frac{1}{\sigma^2}\widehat{\mathbf{w}}(0)\mathbf{x}_1^T \mathbf{I}\mathbf{x}_1 - \frac{1}{\sigma^2}\mathbf{I}\mathbf{x}_1 \mathbf{x}_1^T \widehat{\mathbf{w}}(0) - \frac{1}{\sigma^2} t_1 \mathbf{I}\mathbf{x}_1}{\sigma^2 + \mathbf{x}_1^T \mathbf{I}\mathbf{x}_1} && (38)\\
&= \frac{\widehat{\mathbf{w}}(0) + B\widehat{\mathbf{w}}(0)\mathbf{x}_1^T \mathbf{x}_1 - B\mathbf{x}_1 \mathbf{x}_1^T \widehat{\mathbf{w}}(0) + Bt_1 \mathbf{x}_1}{1 + B\mathbf{x}_1^T \mathbf{x}_1} && (39)\\
\widehat{\mathbf{w}}(1) &= \frac{\widehat{\mathbf{w}}(0) + Bt_1 \mathbf{x}_1}{1 + B\mathbf{x}_1^T \mathbf{x}_1} && (40)
\end{aligned}$$

With $B = \frac{1}{\sigma^2}$. This equation is combined with Equation 33 to solve for the initialization point regardless of data:

$$\frac{A\mathbf{x}_1 t_1}{1 + A\mathbf{x}_1^T \mathbf{x}_1} = \frac{\widehat{\mathbf{w}}(0) + Bt_1 \mathbf{x}_1}{1 + B\mathbf{x}_1^T \mathbf{x}_1} \tag{41}$$

$$\begin{aligned}
\widehat{\mathbf{w}}(0) &= \mathbf{0} && (42)\\
\mathbf{C}_{\mathbf{w}|\mathbf{t}}(0) &= \mathbf{I} && (43)\\
\frac{1}{\sigma^2} &= \mu s^2 && (44)
\end{aligned}$$

1.5 Part E

For this portion of the project, we plot the contour plot of $\widehat{\mathbf{w}}$ at certain points in the sBLR in order to see if it is converging to the correct value. The frames that have been included in the report are the 100th, 250th, 500th, and 750th frames.

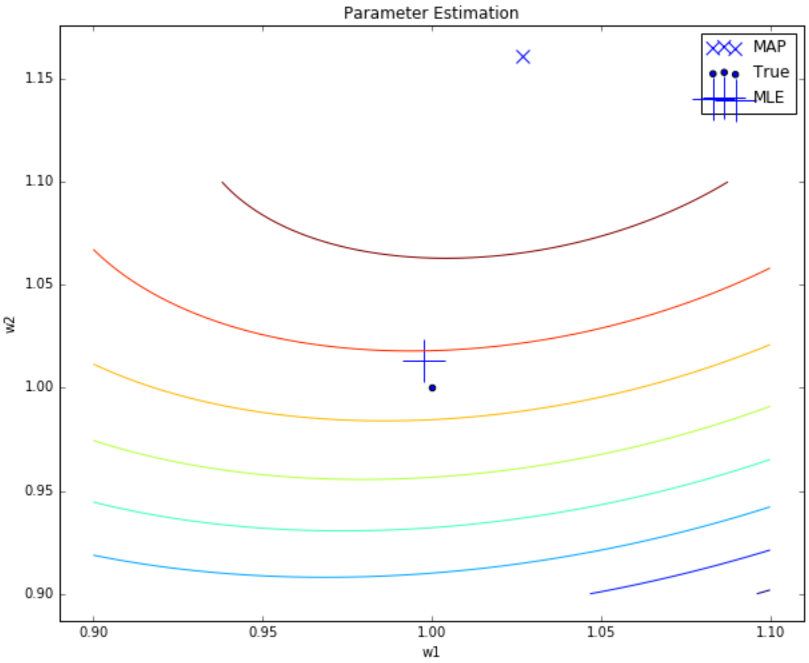


Figure 1: $p(\widehat{\mathbf{w}}|\mathbf{t})$ at frame 100

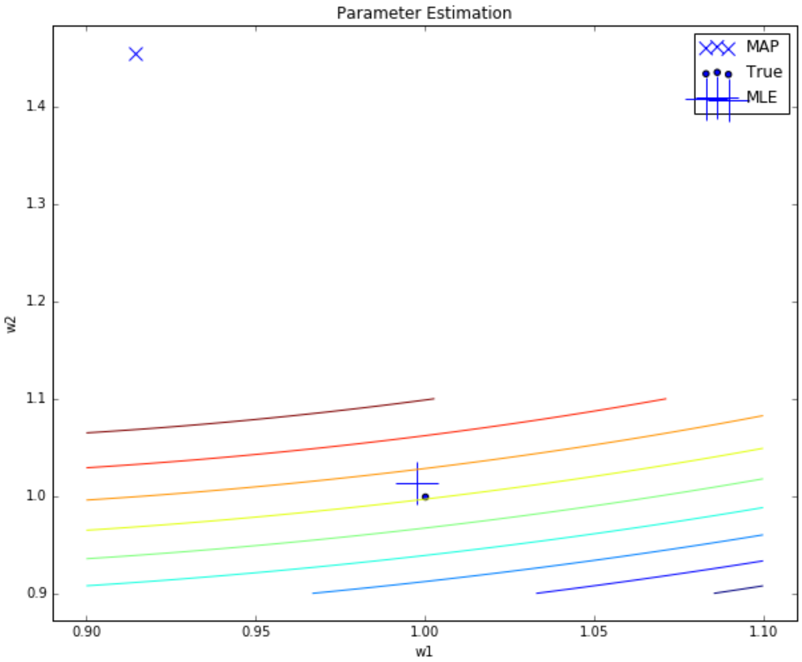


Figure 2: $p(\widehat{\mathbf{w}}|\mathbf{t})$ at frame 250

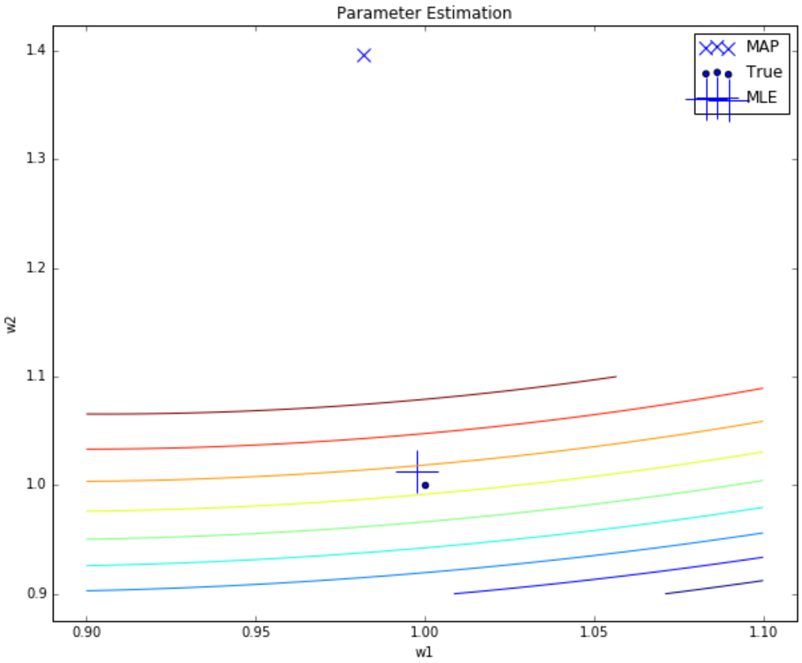


Figure 3: $p(\widehat{\mathbf{w}}|\mathbf{t})$ at frame 500

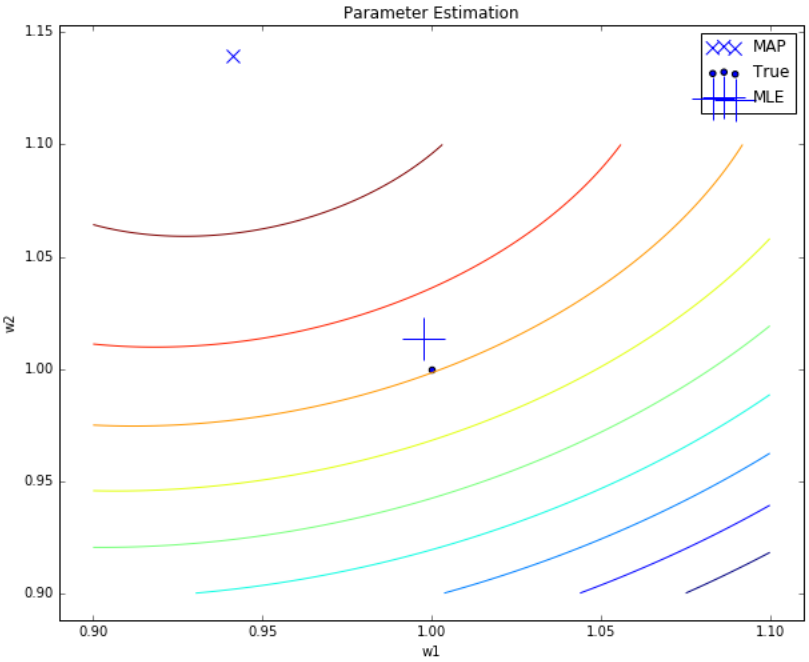


Figure 4: $p(\widehat{\mathbf{w}}|\mathbf{t})$ at frame 750

In each of the above figures, Figures 1, 2, 3, and 4, there are four things being plotted. The true parameter, the MAP prediction, the MLE prediction, and the Gaussian distribution of the MAP prediction. The purpose of both the BLR and sBLR is to converge to the true parameter values to make the best predictive model. The true value occurs at $[1 \;\; 1]^T$ so the analysis is centered around this point.

Both the MAP and MLE parameters do a good job of estimation however the MLE yields a much better job by qualitatively inspecting the graphs. The MLE is not done in a sequential manner so when a large data set is analyzed the potential accuracy increase of the MLE estimate may not be worth the computation load versus the much quick but less accurate MAP estimate.

What is strange is that the contours of the Gaussian distribution do not become more and more peaky as more data is included in the regression. It is to be expected that the covariance of the posterior of our parameters should decrease as more data is collected because we are more confident in our distributions mean. However this does not seem to be the case in the figures. I believe the cause of this is that something is incorrect in either the update equation for the covariance or there is something wrong with the bivariate normal function in matplotlib library of python. Somewhere within the bivariate normal function, the covariance matrix is inverted and will therefore fail if the matrix is singular. My covariance matrix does have singular iterations and therefore I check if the condition number of my covariance is equal to infinity and if it is, I display the last good covariance matrix. The number of good frames that are plotted is 797 which leads me to believe that my sBLR is indeed implemented correctly and that the reason why my Gaussian is not becoming more centered is because of the python scripts that are generating the contours of the posterior parameter distribution.

This is the link to the animation is:

https://drive.google.com/file/d/0Bx00FE5DANNMVVhRWGJSUzB2WEU/view?usp=sharing

1.6 Part F

This portion of the project was to view the Predictive Intervals for the sBLR. The same frames were plotted as in Part E, the 100th, 250th, 500th, and 750th frames.

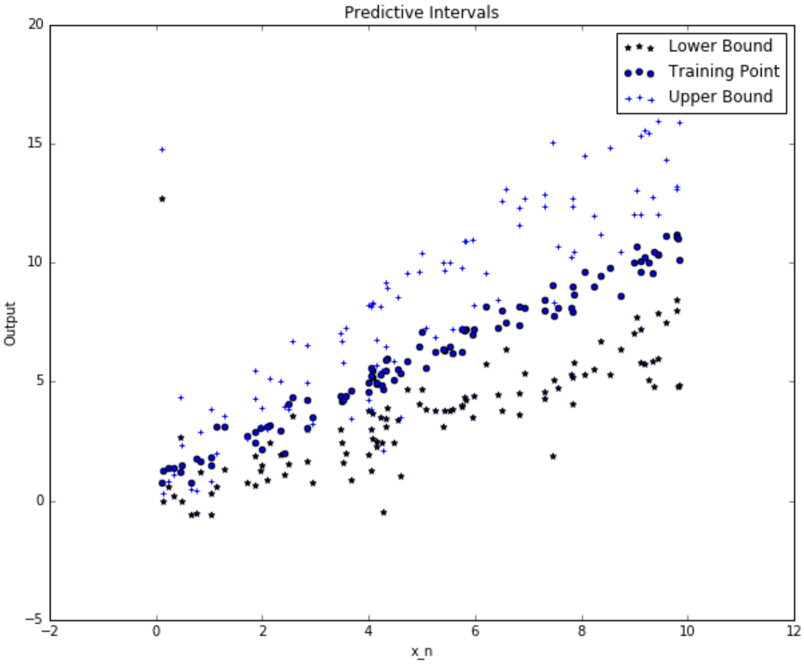


Figure 5: 90% Predictive Interval with 100 data points

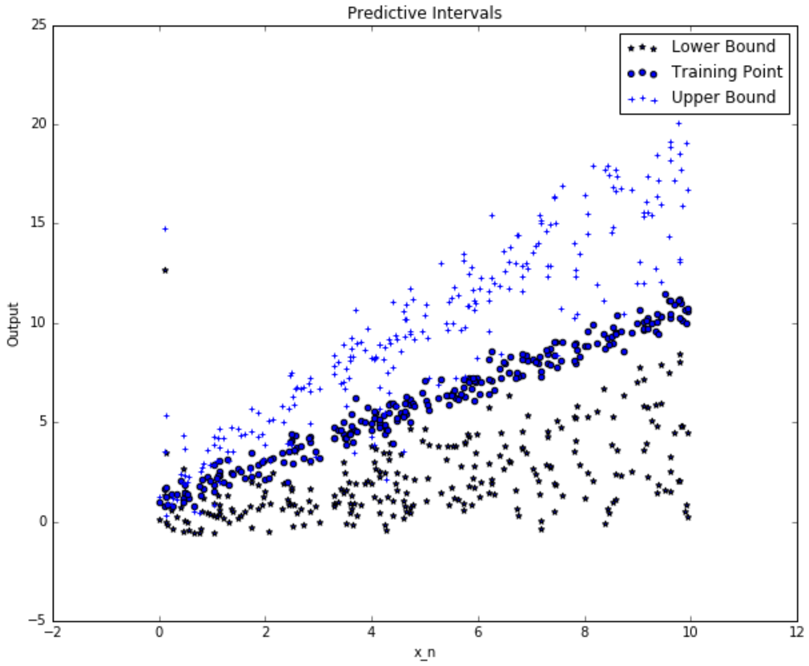


Figure 6: 90% Predictive Interval with 250 data points

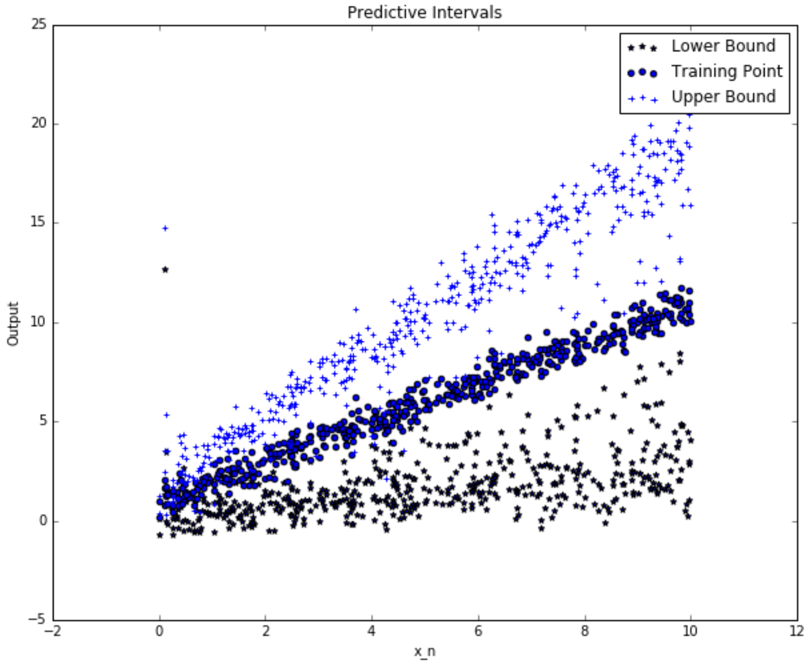


Figure 7: 90% Predictive Interval with 500 data points

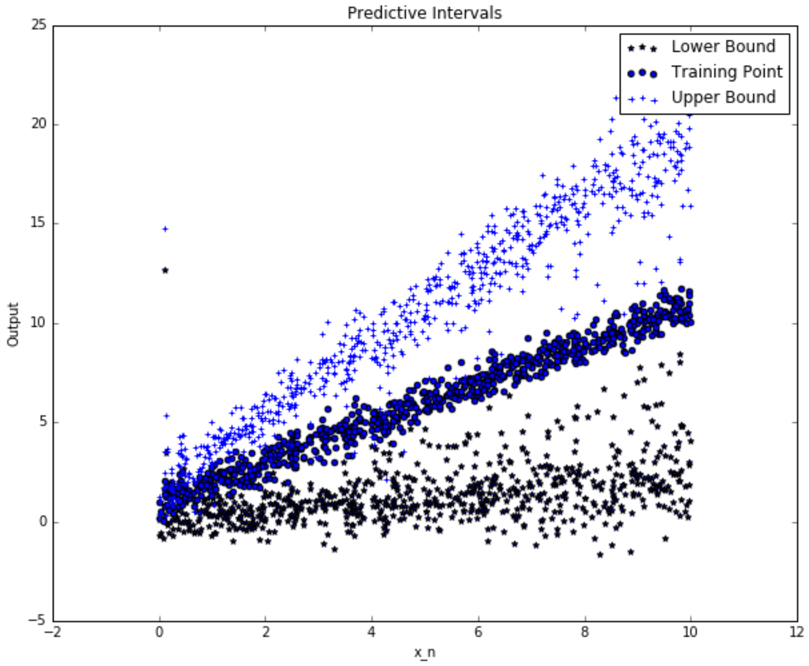


Figure 8: 90% Predictive Interval with 750 data points

For each data point there are points above and below that correspond to the Predictive Interval that we can say with some certainty our prediction will lie within. In order to calculate these intervals I used the following equations:

$$P_l = \widehat{\mathbf{w}}(N)^T\mathbf{x}(N) - \alpha\sqrt{|\sigma^2 + \mathbf{x}(N)^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}(N)|} \tag{45}$$

$$P_h = \widehat{\mathbf{w}}(N)^T\mathbf{x}(N) + \alpha\sqrt{|\sigma^2 + \mathbf{x}(N)^T\mathbf{C}_{\mathbf{w}|\mathbf{t}}(N)\mathbf{x}(N)|} \tag{46}$$

Where α is a hyperparameter that dictates what is the range of confidence that you have and was selected by this equation:

$$\text{Prediction Confidence } = 2 * 100 * (1 - \alpha)\% \tag{47}$$

Which yields a number between 0 and 1. I used $\alpha = .55$ which corresponds to 90% certainty.

As can be seen from the Figures 5, 6, 7, and 8 the 90% Predictive Interval does a very good job of estimating the actual value. There is one outlier at the begining of the dataset and I believe this to be the starting or second data point when we are still very unsure of what our distribution of our MAP estimate is.

The link to the animation is:

https://drive.google.com/file/d/0Bx00FE5DANNMWW1oU08zaHZkQW8/view?usp=sharing

2 Task 2

2.1 Part A

In this section we define a feature vector in order to increase the richness of our regression. However as always, increasing the richness of our model greatly increases the chance of overfitting and therefore cross validation needs to be performed.

In Figure 9 we can see that target values of the training data and of the various orders of feature vectors used in order to perform the regression. From a purely qualitative standpoint, each one of the feature vectors except $p = 12$ does a good job of describing the training output.

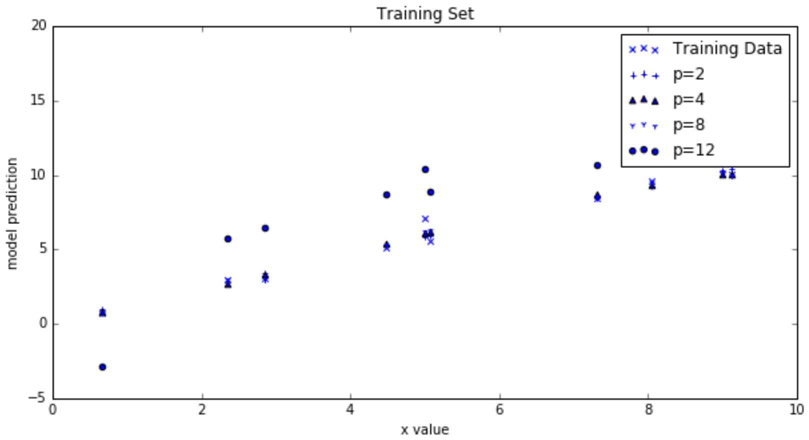


Figure 9: Graphic of the LR model output

2.2 Part B

In order to find the optimal feature vector size, the rest of the dataset was used as a validation set. The MSE error was calculated from the validation set with the trained parameters from the training set and was plotted vs feature vector size as can be seen in Figure 10. The minimum MSE and therefore optimal feature vector size occurs at $p = 2$.

At $p = 1$, our model function looks like $t = w_1x_1 + w_2$ which is a very small transform that does not add much richness to our model and therefore it makes sense that it will not have the smallest MSE and also that the training and validation MSE are close together because it has not had a chance to overfit. However every dimension size above 2, we are greatly increasing the richness of our model and we start to see large overfitting problems above $p = 6$. As we increase our dimensionality up to 6, we see small declines in the training error and increasing validation errors implying that we are seeing overfitting. At $9 > p > 6$ we see a very large increase in the richness of our function and see a large increase in the validation error and continue to see the slight decrease in training error. Then after $p = 9$ we see something very strange, the training error increases dramatically! This is very odd to see that as our model becomes more and more rich, it should do a better and better job of estimating the training data however what does make sense is that the validation greatly increases meaning that the model has overfit the training data very much and we no longer estimate a good model.

Therefore based upon all the above stated reasons $p = 2$ makes sense as a model estimate. We are increasing the richness of our model in order to get a better regression however we are not increasing the richness to much in order to not overfit on our training set.

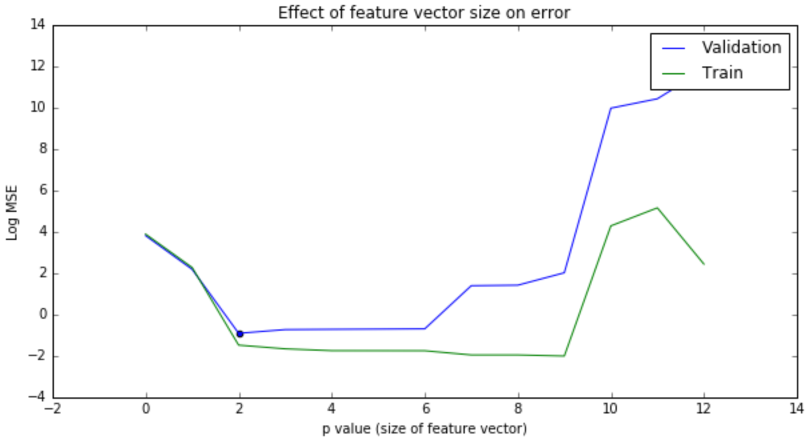


Figure 10: MSE by data transformation dimension

2.3 Part C

Now we apply a regularization term in order to combat the overfitting that we saw in Part B with the large dimension feature vectors. We choose $p = 5$ and then search for the optimal value of the regularization hyperparameter with our validation set. The figure shows that regularization does not greatly help estimate our model and can be seen in Figure 11. The minimizing μ value was 0.004 and the MSE value was 0.049.

The optimal weight of $[0.078 \;\; 1.071 \;\; 0.0029 \;\; 0.0138 \;\; -0.0022 \;\; 0.00008]^T$ that was calculated was slightly similar to the true weight which equals $[1 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$. The first term which corresponds to the addition of the weight is not good because that should be 0 but every other term matches up very nicely.

RR is justified because if we use a rich model function we need a method in order to combat overfitting. Adding in a regularization term will help combat the overfitting and therefore allow us to use more rich functions while not grossly overfitting on the training data. The RR model is justified because it gives a MSE on the validation set of 0.407 and the best LR model gave a MSE of 0.049.

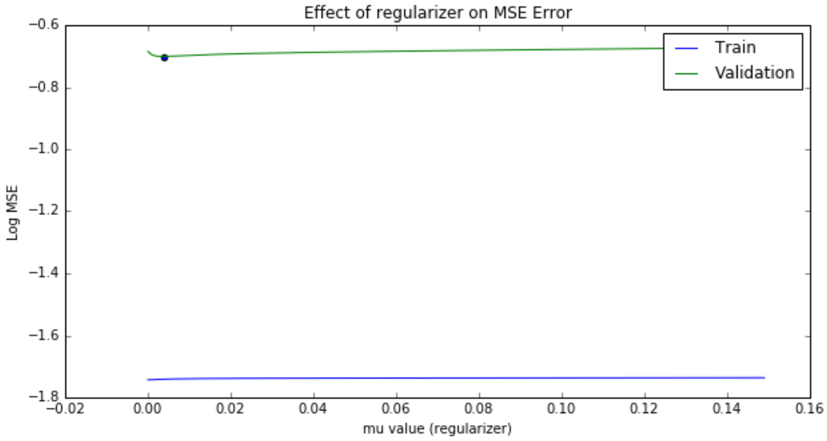


Figure 11: MSE of feature vector RR model with dimension 5

3 Task 3

3.1 Part A

I performed a RR and found that the minimizing regularization hyperparameter μ was 7.1 and the MSE value was 139.78. The results of μ value on MSE can be see in Figure 12

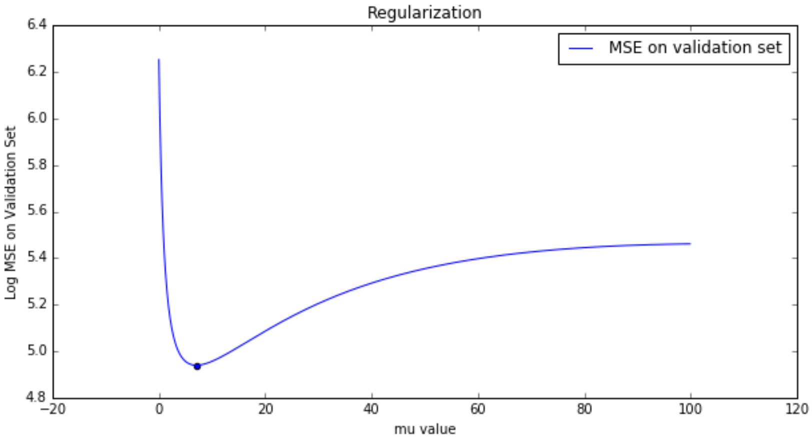


Figure 12: MSE of RR model on the Boston Housing Dataset

3.2 Part B

This portion of the project employed an RBF data transformation. The centers were the data points therefore transforming the training data into a square matrix. I ran into many issues finding good values of the hyperparameter γ that is within the transformation kernel because the exponential so quickly will crash. In order to find a good starting point for γ I found the median distance between all the training vectors and started my search there. When I searched below the start point I found only noise so I started to search above as well. The transformation runs in N^2 time and therefore was very time consuming so I used very small search intervals and had to search many times before I found the window seen in Figure 13. When I searched above the median value, I kept finding optimal values at the last searched value and therefore kept moving my window over and eventually I found the minimizing γ at 0.00106 and the MSE value was 258.73.

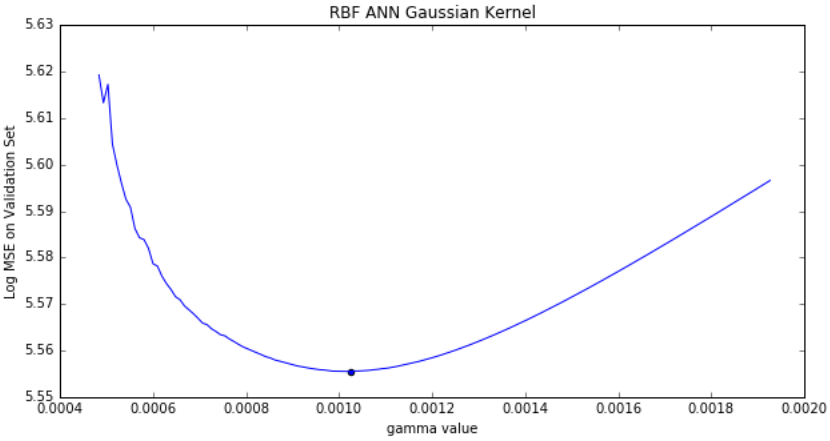
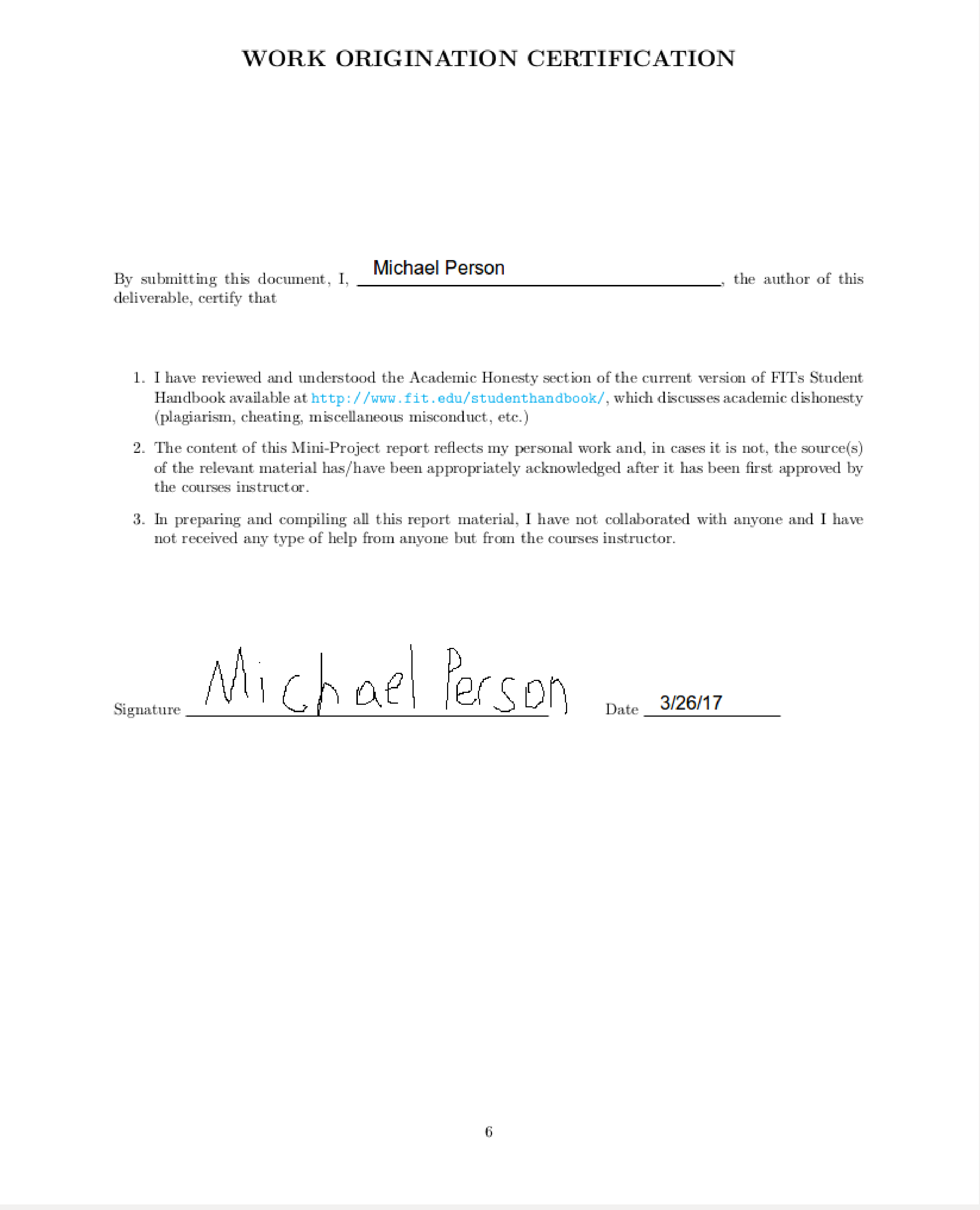


Figure 13: MSE of RBF model on the Boston Housing Dataset

3.3 Part C

When the champion RR and RBF model were evaluated on the test set, I reached a MSE of 52.39 and 236.56 making the RR the overall champion model for this dataset. The RR model was then compared against the supplied LR model which scored an MSE of 4505.87 therefore making the RR model the overall champion. The LR model represented an average of all the attributes in the given data point in order to make a prediction.

Given that the MSE of the overall champion was 52.39 was can conclude that for any given prediction, we will be roughly 7 away from the true value. Our targets were prices of houses in thousands of dollars meaning that we are able to predict the price of a house within 7 thousand dollars using only our model. Therefore we can say with increasing certainty that our model is accurate as the price of the house increases. If we use the model to analyze cheap houses then the 7 thousand dollars will yield greater error but if we analyze a house that costs hundreds of thousands of dollars if not millions of dollars, we can extremely accurately predict what the value of the house will be. Therefore since most houses are in the hundreds of thousands or greater regime, our model does an excellent job of predicting housing prices.



WORK ORIGINATION CERTIFICATION

By submitting this document, I, Michael Person, the author of this deliverable, certify that

1. I have reviewed and understood the Academic Honesty section of the current version of FITs Student Handbook available at http://www.fit.edu/studenthandbook/, which discusses academic dishonesty (plagiarism, cheating, miscellaneous misconduct, etc.)
2. The content of this Mini-Project report reflects my personal work and, in cases it is not, the source(s) of the relevant material has/have been appropriately acknowledged after it has been first approved by the courses instructor.
3. In preparing and compiling all this report material, I have not collaborated with anyone and I have not received any type of help from anyone but from the courses instructor.

Signature Michael Person Date 3/26/17

References

[1] Sherman-morrison formula. URL: https://en.wikipedia.org/wiki/Sherman%E2%80%93Morrison_formula.