Homework 1

CSCI 5525: Machine Learning

Due on September 26th 11am (before class)

Please type in your info:

- **Name**: Graham Droge
- **Student ID**: 5478165
- **Email**: droge010@umn.edu
- **Collaborators, and on which problems:** None

Homework Policy. (1) You are encouraged to collaborate with your classmates on homework problems, but each person must write up the final solutions individually. You need to fill in above to specify which problems were a collaborative effort and with whom. (2) Regarding online resources, you should **not**:

- Google around for solutions to homework problems,
- Ask for help on online.
- Look up things/post on sites like Quora, StackExchange, etc.

Submission. Submit a PDF using this LaTeX template for written assignment part and submit Python jupyter or Colab python notebooks (.ipynb) for all programming part. You should upload all the files on Canvas.

Written Assignment

Instruction. For each problem, you are required to write down a full mathematical proof to establish the claim.

Problem 1. Two helpful matrices.

Let us first recall the notations in linear regression. The design matrix and the response vector are are defined as:

$$A = \begin{bmatrix} \leftarrow x_1^\intercal \rightarrow \\ \vdots \\ \leftarrow x_n^\intercal \rightarrow \end{bmatrix} \qquad \mathbf{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

For this problem, we will assume the covariance matrix $A^\intercal A$ is invertible, and so $(A^\intercal A)^{-1}$ is well-defined (**Clearly mention the properties of matrix operations used while solving**).

Problem 1.1. The residual matrix. For any weight vector $\mathbf{w}$, let us define the vector of least squares residuals as

$$e = \mathbf{b} - A\mathbf{w}$$

Now if $\mathbf{w}$ is the least square solution given by $\mathbf{w} = (A^\intercal A)^{-1}A^\intercal\mathbf{b}$, we can rewrite e as

$$e = \mathbf{b} - A(A^\intercal A)^{-1}A^\intercal\mathbf{b} = \left(I - A(A^\intercal A)^{-1}A^\intercal\right)\mathbf{b}$$

Now let $M = \left(I - A(A^\intercal A)^{-1}A^\intercal\right)$. Show that

- M is symmetric (i.e. $M = M^\intercal$). (**2 points**)
- M is idempotent (i.e. $M^2 = M$). (**2 points**)
- $MA = 0$. (**1 point**)

Your answer. **1)** M is symmetric

The first thing we do to solve this is make the observation that subtracting $A(A^\intercal A)^{-1}A^\intercal$ from the identity matrix does not alter the symmetry. This makes sense since the identity matrix is only non-zero along the trace and the transpose does not affect the diagonal elements. Now we can just test for symmetry for the second term i.e show the following

$$A(A^\intercal A)^{-1}A^\intercal = (A(A^\intercal A)^{-1}A^\intercal)^\intercal \quad (1)$$

Working with the above expression our next step is to bring the transpose into the parentheses for the right hand term. To do this we use the following property of matrix transposes where B, C, D are matrices

$$(BCD)^\intercal = D^\intercal C^\intercal B^\intercal$$

The property says to filter the transpose through the parentheses you take the transpose of individual terms and reverse the ordering. So now lets make some definitions to take advantage of this property

$$B = A \quad C = (A^\intercal A)^{-1} \quad D = A^\intercal$$

Thus

$$(A(A^\intercal A)^{-1}A^\intercal)^\intercal = A((A^\intercal A)^{-1})^\intercal A^\intercal \quad (2)$$

The interesting term that we want to analyze now is the $((A^\intercal A)^{-1})^\intercal$ term. Using more properties of the transpose we can write

$$((A^\intercal A)^{-1})^\intercal = ((A^\intercal A)^\intercal)^{-1} \quad (A^\intercal A)^\intercal = A^\intercal(A^\intercal)^\intercal = A^\intercal A$$

Thus we can write

$$((A^\intercal A)^{-1})^\intercal = (A^\intercal A)^{-1} \quad (3)$$

Plugging (3) into (2) we obtain

$$(A(A^\intercal A)^{-1}A^\intercal)^\intercal = A(A^\intercal A)^{-1}A^\intercal \quad (4)$$

And comparing (4) with (1) and using the initial observation that the identity matrix does not affect the symmetry we can see that the equality $M = M^\intercal$ does in fact hold

2) M is idempotent

Start by defining $M^* = A(A^\intercal A)^{-1}A^\intercal$. Thus

$$M^2 = M$$

$$(I - M^*)(I - M^*) = I - 2M^* + M^*M^* \quad (1)$$

Now lets look specifically at the term M^*M^*

$$M^*M^* = (A(A^\intercal A)^{-1}A^\intercal)(A(A^\intercal A)^{-1}A^\intercal) = A(A^\intercal A)^{-1}[A^\intercal A](A^\intercal A)^{-1}A^\intercal$$

I put the two terms in parentheses into one term with the [ ] being used to provide separation from other terms. We can see that the term $A^\intercal A(A^\intercal A)^{-1} = I$ so we can write

$$M^*M^* = A(A^\intercal A)^{-1}A^\intercal = M^* \quad (2)$$

Lets plug (2) back into (1)

$$M^2 = I - 2M^* + M^* = I - M^* = M$$

Thus we proved that $M^2 = M$

3) $MA = 0$

This one can be proved in a rather straightforward way

$$MA = (I - A(A^\intercal A)^{-1}A^\intercal)A = A - A(A^\intercal A)^{-1}A^\intercal A$$

From here we notice that $(A^\intercal A)^{-1}A^\intercal A = I$ Thus

$$MA = A - A = 0$$

And we have proven $MA = 0$

Problem 1.2. The hat matrix. Using the residual maker, we can derive another matrix, the hat matrix or projection matrix $P = I - M = A(A^\intercal A)^{-1}A^\intercal$. Note that the predicted value by the least squares solution is given by $P\mathbf{b}$. Show that

- P is symmetric. (**1 point**)
- P is idempotent. (**1 point**)

Your answer. **1)** P is symmetric

We have actually already proved both of these in the previous problem so we'll use the same method as before.

$$P = A(A^\intercal A)^{-1}A^\intercal$$

So we need to show

$$A(A^\intercal A)^{-1}A^\intercal = (A(A^\intercal A)^{-1}A^\intercal)^\intercal$$

Using the transpose properties as in problem one we can write it as

$$A(A^\intercal A)^{-1}A^\intercal = (A(A^\intercal A)^{-1}A^\intercal)^\intercal = A((A^\intercal A)^{-1})^\intercal A^\intercal \quad (1)$$

Then we use the following properties

$$((A^\intercal A)^{-1})^\intercal = ((A^\intercal A)^\intercal)^{-1} \quad (2)$$

$$(A^\intercal A)^\intercal = A^\intercal (A^\intercal)^\intercal = A^\intercal A \quad (3)$$

Plugging (3) into (2) and then pluggin that into (1) we obtain

$$A(A^\intercal A)^{-1}A^\intercal = A(A^\intercal A)^{-1}A^\intercal$$

So it is symmetric.

2) P is idempotent

We will use the same method that was used in the previous problem

$$P^2 = (A(A^\intercal A)^{-1}A^\intercal)(A(A^\intercal A)^{-1}A^\intercal) = A(A^\intercal A)^{-1}[A^\intercal A](A^\intercal A)^{-1}A^\intercal$$

We observe that the previous equation can be simplified by applying the relation $I = A^\intercal A(A^\intercal A)^{-1}$
So we are left with

$$P^2 = A(A^\intercal A)^{-1}A^\intercal = P$$

Problem 2. Gradient of conditional log-likelihood.

For any $a \in \mathbb{R}$, let $\sigma(a) = \frac{1}{1+\exp(-a)}$. For each example $(x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$, the conditional log-likelihood of logistic regression is

$$\ell(y_i \mid x_i, \mathbf{w}) = y_i \ln(\sigma(\mathbf{w}^\intercal x_i)) + (1 - y_i) \ln(\sigma(-\mathbf{w}^\intercal x_i))$$

Derive the gradient of $\ell(y_i \mid x_i, \mathbf{w})$ with respect to w_j (i.e. the j-th coordinate of $\mathbf{w}$) by following the following steps (**Clearly mention the properties of derivatives used while solving**).

- Derive $\frac{\partial}{\partial a}\sigma(a)$. (**2 points**)
- Derive $\frac{\partial}{\partial w_j}\sigma(\mathbf{w}^\intercal x_i)$. (**1 point**)
- Derive $\frac{\partial}{\partial w_j}\ln\sigma(\mathbf{w}^\intercal x_i)$. (**2 points**)
- Derive $\frac{\partial}{\partial w_j}\ln\sigma(-\mathbf{w}^\intercal x_i)$. (**1 point**)
- Derive $\frac{\partial}{\partial w_j}\ell(y_i \mid x_i, \mathbf{w})$. (**2 points**)

Your answer. **1)** Derive $\frac{\partial}{\partial a}\sigma(a)$

Now we take advantage of the chain rule to differentiate the function $\frac{1}{1+\exp(-a)}$

$$\frac{\partial}{\partial a}(1 + \exp(-a))^{-1} = -(1 + \exp\{-a\})^{-2}(-\exp\{-a\}) = \frac{\exp\{-a\}}{(1 + \exp\{-a\})^2}$$

2) Derive $\frac{\partial}{\partial w_j}\sigma(\mathbf{w}^\intercal x_i)$

We will use the same chain rule to differentiate w.r.t a specific weight $\frac{\partial \sigma(\mathbf{w}^\intercal x_i)}{\partial w_j} = \frac{\partial \sigma(a)}{\partial a}\frac{\partial a}{\partial w_j}$. We can use the previous derivations result but now take $\frac{\partial a}{\partial w_j}$

$$\frac{\partial}{\partial w_j}(1+\exp(-\mathbf{w}^\intercal x_i))^{-1} = -(1+\exp\{-\mathbf{w}^\intercal x_i\})^{-2}(-\frac{\partial \mathbf{w}^\intercal x_i}{\partial w_j}\exp\{-\mathbf{w}^\intercal x_i\}) = \frac{x_i \exp\{-\mathbf{w}^\intercal x_i\}}{(1+\exp\{-\mathbf{w}^\intercal x_i\})^2}$$

3) Derive $\frac{\partial}{\partial w_j}\ln\sigma(\mathbf{w}^\intercal x_i)$

Once again we use the chain rule but this time our first partial derivative acts on the natural logarithm as opposed to $\frac{1}{1+\exp\{a\}}$

$$\frac{\partial}{\partial x}\ln(a) = \frac{1}{a}\frac{\partial a}{\partial x}$$

$$\frac{\partial}{\partial w_j}\ln(\sigma(\mathbf{w}^\intercal x_i)) = \frac{1}{\sigma(\mathbf{w}^\intercal x_i)}\frac{\partial}{\partial w_j}\sigma(\mathbf{w}^\intercal x_i)$$

Now we have already calculated the rightmost term in the previous equation so we can then write

$$\frac{\partial}{\partial w_j}\ln(\sigma(\mathbf{w}^\intercal x_i)) = -(1+\exp\{-\mathbf{w}^\intercal x_i\})\frac{x_i \exp\{-\mathbf{w}^\intercal x_i\}}{(1+\exp\{-\mathbf{w}^\intercal x_i\})^2}$$

$$\frac{\partial}{\partial w_j}\ln(\sigma(\mathbf{w}^\intercal x_i)) = -\frac{x_i \exp\{-\mathbf{w}^\intercal x_i\}}{(1+\exp\{-\mathbf{w}^\intercal x_i\})}$$

and we're done

4) 3) Derive $\frac{\partial}{\partial w_j}\ln\sigma(-\mathbf{w}^\intercal x_i)$

This can be proved simply since it is just the previous derivation with an added negative term that needs to propagate through the chain rule terms

$$\frac{\partial}{\partial w_j}\ln(\sigma(-\mathbf{w}^\intercal x_i)) = (1+\exp\{\mathbf{w}^\intercal x_i\})\frac{x_i \exp\{\mathbf{w}^\intercal x_i\}}{(1+\exp\{\mathbf{w}^\intercal x_i\})^2}$$

and we get

$$\frac{\partial}{\partial w_j}\ln(\sigma(\mathbf{w}^\intercal x_i)) = \frac{x_i \exp\{\mathbf{w}^\intercal x_i\}}{(1+\exp\{\mathbf{w}^\intercal x_i\})}$$

An important relationship that we notice is the

$$\frac{\partial}{\partial w_j}\ln(\sigma(\mathbf{w}^\intercal x_i)) = \frac{x_i \exp\{\mathbf{w}^\intercal x_i\}}{(1+\exp\{\mathbf{w}^\intercal x_i\})} = 1 - \frac{\partial}{\partial w_j}\ln(\sigma(-\mathbf{w}^\intercal x_i))$$

5) Derive $\frac{\partial}{\partial w_j} l(y_i|x_i,\mathbf{w}) = y_i ln(\sigma(\mathbf{w}^\intercal x_i)) + (1-y_i) ln(\sigma(-\mathbf{w}^\intercal x_i))$

So we already derived all of the needed terms to complete this derivation. Since y_i and $(1-y_i)$ are constant over the partial derivative we can factor them out and use the property of linearity of derivatives to write

$$\frac{\partial}{\partial w_j} l(y_i|x_i, \mathbf{w}) = y_i \frac{\partial}{\partial w_j} ln(\sigma(\mathbf{w}^\intercal x_i)) + (1-y_i)\frac{\partial}{\partial w_j} ln(\sigma(-\mathbf{w}^\intercal x_i))$$

Now all that is left is to apply the previous derivations for $\frac{\partial}{\partial w_j} ln(\sigma(\mathbf{w}^\intercal x_i))$ and $\frac{\partial}{\partial w_j} ln(\sigma(-\mathbf{w}^\intercal x_i))$

$$\frac{\partial}{\partial w_j} l(y_i|x_i, \mathbf{w}) = -y_i \frac{x_i \exp\{-\mathbf{w}^\intercal x_i\}}{(1+\exp\{-\mathbf{w}^\intercal x_i\})} + (1-y_i)\frac{x_i \exp\{\mathbf{w}^\intercal x_i\}}{(1+\exp\{\mathbf{w}^\intercal x_i\})}$$

Now if we were to put this in a simpler and vectorized form it would be of the form

$$\frac{\partial}{\partial w_j} l(y|x, \mathbf{w}) = \frac{1}{n} x^\intercal(\sigma(\mathbf{w}^\intercal x) - y)$$

which we will take advantage of in the programming part of this assignment

Problem 3. Derivation of Ridge Regression Solution.

Recall that in class we claim that the solution to ridge regression ERM:

$$\min_{\mathbf{w}} \ (\|A\mathbf{w} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{w}\|_2^2)$$

is $\mathbf{w}^* = (A^\intercal A + \lambda I)^{-1} A^\intercal \mathbf{b}$. Now provide a proof. (**5 points**)
(Hint: recall that $\nabla F(\mathbf{w}) = \mathbf{0}$ is a sufficient condition for $\mathbf{w}$ to be a minimizer of any convex function F.) (**Clearly mention the properties of matrix calculus used while solving**)

Your answer. .

To start we first compute the gradient of the ridge regression ERM equation given above w.r.t $\mathbf{w}$. To do this we compute the gradient of the norm terms separately and add the results

$$\boldsymbol{\nabla}\Big(\lambda\|\mathbf{w}\|_2^2\Big) = \lambda\boldsymbol{\nabla}\left(\sqrt{w_1^2 + w_2^2 + ...}^{\,2}\right) = \lambda\boldsymbol{\nabla}\big(w_1^2 + w_2^2 + ...\big)$$

The gradient of the previous equation is a vector that contains the partial derivatives for each element Thus

$$\lambda\boldsymbol{\nabla}\big(w_1^2 + w_2^2 + ...\big) = 2\lambda\mathbf{w}$$

Now we compute the gradient of the first term

$$\boldsymbol{\nabla}\|A\mathbf{w} - \mathbf{b}\|_2^2 = \boldsymbol{\nabla}[(A\mathbf{w} - \mathbf{b})^\intercal(A\mathbf{w} - \mathbf{b})]$$

Next we filter the transpose into the terms in the parentheses to get

$$(A\mathbf{w} - \mathbf{b})^\intercal = \mathbf{w}^\intercal A^\intercal - \mathbf{b}^\intercal$$

new plug this back into the previous equation to get

$$\|A\mathbf{w} - \mathbf{b}\|_2^2 = (\mathbf{w}^\intercal A^\intercal - \mathbf{b}^\intercal)(A\mathbf{w} - \mathbf{b}) = \mathbf{w}^\intercal A^\intercal A\mathbf{w} - \mathbf{w}^\intercal A^\intercal \mathbf{b} - \mathbf{b}^\intercal A\mathbf{w} + \mathbf{b}^\intercal\mathbf{b}$$

$$\boldsymbol{\nabla}(\mathbf{w}^\intercal A^\intercal A\mathbf{w} - \mathbf{w}^\intercal A^\intercal \mathbf{b} - \mathbf{b}^\intercal A\mathbf{w} + \mathbf{b}^\intercal \mathbf{b}) = 2A^\intercal A\mathbf{w} - A^\intercal \mathbf{b} - \mathbf{b}^\intercal A$$

We can simplify this further by noticing that $A^\intercal \mathbf{b}$ and $\mathbf{b}^\intercal A$ are equal. The order in which the matrix operations is reversed but the same row by column multiplications are still taking place thus they are equal i.e they are symmetric. So we can simplify to

$$\boldsymbol{\nabla}(\mathbf{w}^\intercal A^\intercal A\mathbf{w} - \mathbf{w}^\intercal A^\intercal \mathbf{b} - \mathbf{b}^\intercal A\mathbf{w} + \mathbf{b}^\intercal \mathbf{b}) = 2A^\intercal A\mathbf{w} - 2A^\intercal \mathbf{b}$$

Now we put the two gradient terms into one equation and setting it equal to zero since that is a sufficient condition for $\mathbf{w}$ to be a minimizer of any convex function

$$\boldsymbol{\nabla}\left(\|A\mathbf{w} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{w}\|_2^2\right) = 0 = 2A^\intercal A\mathbf{w} - 2A^\intercal \mathbf{b} + 2\lambda\mathbf{w}$$

Rearrange the equation to get the final $\mathbf{w}^*$

$$2A^\intercal \mathbf{b} = 2A^\intercal A\mathbf{w} + 2\lambda\mathbf{w}$$

$$A^\intercal \mathbf{b} = (A^\intercal A + \lambda\mathbf{I})\mathbf{w}$$

$$\mathbf{w}^* = (A^\intercal A + \lambda\mathbf{I})^{-1}A^\intercal \mathbf{b}$$

Programming Assignment

Problem 1. Ridge Regression

b)

The assignment calls for us to solve the ridge regression without gradient descent, thus my approach is to find the weight vector wthat produces a gradient of the cost function equal to 0 i.e

$$w^* = argmin\|y - Xw\|^2 + \lambda\|w\|^2$$

$$\nabla C_w = 0 \quad w^* = (\lambda I + X^\intercal X)^{-1} X^\intercal y$$

Since ridge regression was designed to handle singular feature matrices we don't have to worry about singular matrices and we can used the closed form solution for the minimal weight vector. I build weight vectors over the range of lambda's 0-100 and the combination of these weight vectors form the model. Once we calculate the model we calculate the prediction on the next set over the range of lambda's and finally calculate the RMSE the same way so we get RMSE values over the entire range of lambda values. This is done for 5-folds and the averages are compared as well as the individual RMSE values over the range of lambda's.

c)

After running our algorithm we obtain RMSE values for each choice of λ as well as averages for the RMSE values across each of the 5 folds. The average RMSE value across each fold is

$$Fold1 = 6.58, Fold2 = 4.94, Fold3 = 4.96, Fold4 = 4.78, Fold5 = 4.79$$

Thus average across the fold averages is calculated to be 5.202. Along with the the plot in figure 1 shows the average RMSE value across all of the folds for a given λ. Figure 2 shows the RMSE values across the range of λ's for each of the folds for a given iteration.

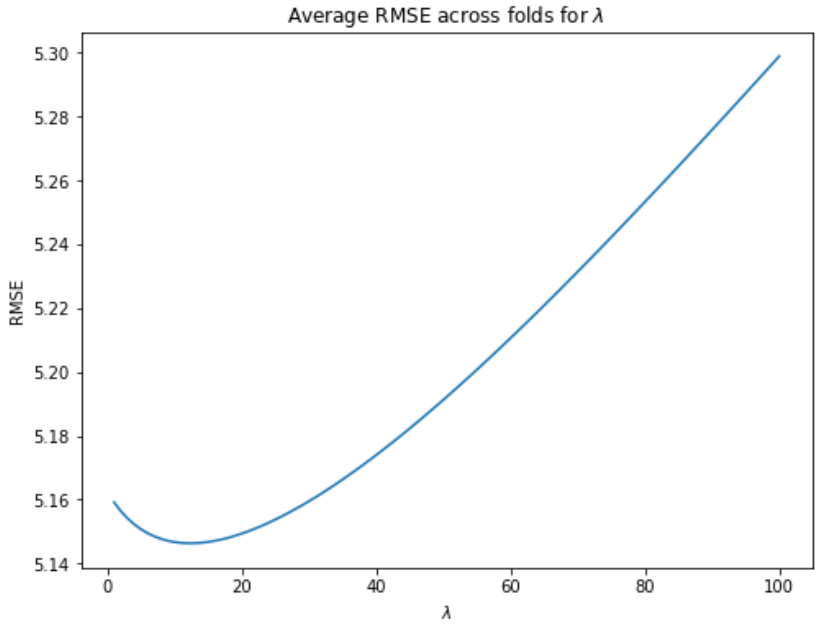

Figure 1: Average RMSE value across the folds for a given λ

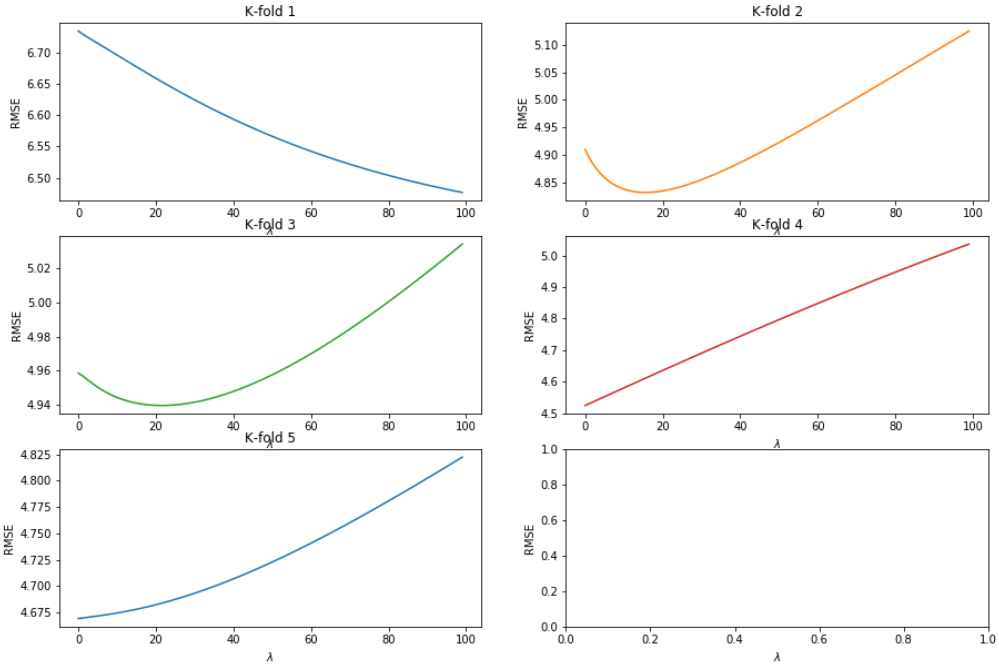

Figure 2: RMSE values for each fold for a given λ

Problem 2. Logistic Regression

b)

For this problem we solve the logistic regression problem which deals with classifying elements into a specific category. To do this we take advantage of the sigmoid function, which has the characteristic of converting a number into a probability. The loss function for logistic regression is given below, as well as the gradient for the loss function that we generalized from when we calculated the partial derivatives with respect to a weight for individual features and labels. Using these equations we train the model by first initializing the weight vector with random values, then compute a guess by taking the sigmoid function of our weight vector multiplied by our training set. Next we compute the gradient by using the previously computed probability guesses and the equation mentioned below. Lastly we compute the next set of weights by subtracting our learning rate multiplied by the grad from our last set of weights. We do this over 50,000 iterations to allow the algorithm to find a minimum location.

$$\ell(y_i \mid x_i, \mathbf{w}) = y_i \ln(\sigma(\mathbf{w}^\intercal x_i)) + (1 - y_i) \ln(\sigma(-\mathbf{w}^\intercal x_i))$$

$$\frac{\partial}{\partial w_j} l(y|x, \mathbf{w}) = \frac{1}{n} x^\intercal (\sigma(\mathbf{w}^\intercal x) - y)$$

c)

Using a learning rate of 0.01 the error rate for each of the validation/training folds were

$$Fold1 = 0, Fold2 = 0, Fold3 = 0, Fold4 = 0, Fold5 = .033$$

with the average across all of them being .006. The learing rate selection of .01 was chosen after performing some hyperparameter testing. Very small values of the learning rate didnt allow enough progression towards the minimum so the error rate was higher. When the learning rate was high the next step in the descent would overshoot the minimum so settling at the minimum became harder so it was higher as well. The learning rate of .01 was a good compromise between the two and provided a good RMSE value. A plot of the RMSE over the learning rates is shown.

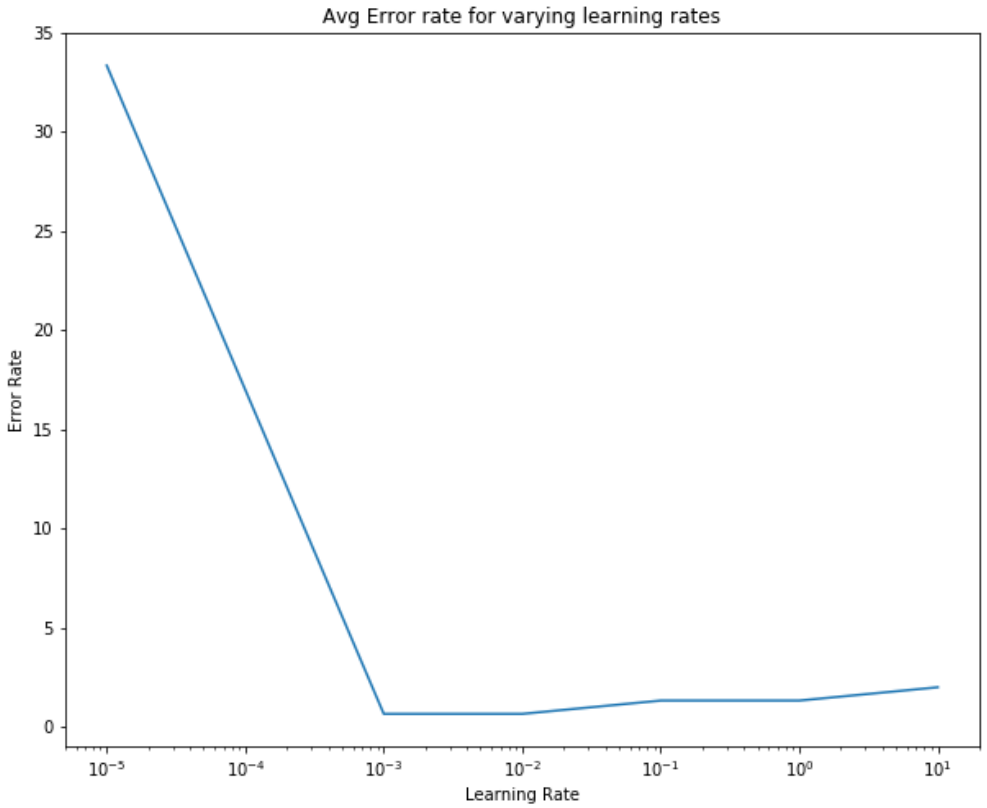

Figure 3: RMSE over various learning rates