Problem 5 [Ridge regression] 20 points

In this problem, you will derive the optimal parameters for ridge regression and implement a ridge regression model. In ridge regression, the loss function includes a regularization term:

$$J(\theta) = \sum_{i=1}^{n}[h_\theta(x^{(i)}) - y^{(i)}]^2 + \lambda \sum_{j=1}^{d} \theta_j^2$$

(a) Write the derivation of the closed form solution for parameter θ that minimizes the loss function $J(\theta)$ in ridge regression.

(b) Modify your linear regression implementation to handle ridge regression. Compare the results of linear regression and ridge regression on the dataset. Take several values of the regularization parameter λ and output the MSE, RSE, and R^2 metrics. Which model performs better? Interpret the results in your own words.

Derivation of the closed form solution of Ridge Regression
Begin with the definition of the loss function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}[h_\theta(x^{(i)}) - y^{(i)}]^2 + \frac{1}{2}\lambda \sum_{j=1}^{d} \theta_j^2$$

where d is the dimensions of the training data
$h_\theta(x) = \sum_{m=0}^{d} \theta_m x_m$
$x^{(i)}$ denotes the ith datapoint of the training dataset
We resolve to find the minima of a convex function defined by the cost function $J(\theta)$
Generally the approach will be to:

1. Find the gradient of the function $J(\theta)$
2. Set the gradient to zero
3. Solve for the θ

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left[h_\theta(x^{(i)}) - y^{(i)}\right]^2 + \frac{1}{2}\lambda \sum_{j=1}^{d} \theta_j^2$$

$$\begin{aligned}\frac{\partial J\theta}{\partial \theta_l} = \frac{\partial}{\partial \theta_l}\frac{1}{2}\left[\sum_{i=1}^{n}\left[\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_l x_l^{(i)}\right] - y^{(i)}\right]^2 \\ + \frac{\partial}{\partial \theta_l}\left[\frac{\lambda_l}{2}\sum_{j=1}^{d}\theta_j\right]\end{aligned}$$

For readability we now take the left half of the sum to compute

$$\frac{\partial}{\partial \theta_l}\frac{1}{2}\left[\sum_{i=1}^{n}\left[\theta_0+\theta_1 x_1^{(i)}+\theta_2 x_2^{(i)}+\cdots+\theta_l x_l^{(i)}\right]-y^{(i)}\right]^2$$

$$=\frac{1}{2}\left[\sum_{i=1}^{n}\left[\frac{\partial}{\partial \theta_l}(\theta_0)+\frac{\partial}{\partial \theta_l}\left(\theta_1 x_1^{(i)}\right)+\frac{\partial}{\partial \theta_l}\left(\theta_2 x_2^{(i)}\right)+\cdots+\frac{\partial}{\partial \theta_l}\left(\theta_l x_l^{(i)}\right)+\cdots+\frac{\partial}{\partial \theta_l}\left(\theta_k x_k^{(i)}\right)\right]-\frac{\partial}{\partial \theta_l}y^{(i)}\right]$$

$$\cdot 2\cdot\left[\sum_{i=0}^{n}\left[\sum_{k=0}^{d}\left[\theta_k x_k^{(i)}\right]-y^{(i)}\right]\right]$$

$$=x_l^{(i)}\sum_{i=1}^{n}\left[\sum_{k=0}^{d}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]\right]$$

For $0 < l \leq k$

Now to reintroduce the right hand sum

$$\frac{\partial}{\partial \theta_l}\left[\frac{\lambda_l}{2}\sum_{j=1}^{d}\theta_j\right]$$

$$=\frac{\lambda_l}{2}\cdot 2\theta_l \qquad \text{(Given that } l \in [1,d]\text{)}$$

$$=\lambda_l \theta_l$$

Combining equations again

$$\frac{\partial J(\theta)}{\partial}x_l^{(i)}\sum_{i=1}^{n}\left[\sum_{k=0}^{d}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]\right]+\lambda_l\theta_l=0$$

$$=x_l^{(i)}\sum_{i=1}^{n}\left[h_\theta(x^{(i)})-y^{(i)}\right]+\lambda_l\theta_l=0$$

$$=\sum_{i=1}^{n}x_l^{(i)}\left[\sum_{k=0}^{l-1}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]+\left(\theta_l x_l^{(i)}-y^{(i)}\right)+\sum_{k=l+1}^{d}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]\right]+\lambda_l\theta_l$$

$$=\sum_{i=1}^{n}\left(x_l^{(i)}\right)^2\theta_l\left[\sum_{k=0}^{l-1}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]-y^{(i)}+\sum_{k=l+1}^{d}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]\right]+\lambda_l\theta_l$$

$$=\theta_l\left(\sum_{i=1}^{n}\left(x_l^{(i)}\right)^2\left[\sum_{k=0}^{l-1}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]-y^{(i)}+\sum_{k=l+1}^{d}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]\right]\right)+\lambda_l$$

$$\theta_l=-\lambda_l\left(\sum_{i=1}^{n}\left(x_l^{(i)}\right)^2\left[\sum_{k=0}^{l-1}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]+\sum_{k=l+1}^{d}\left[\left(\theta_k x_k^{(i)}\right)-y^{(i)}\right]-y^{(i)}\right]\right)^{-1}$$

From this point we could calculate all other values of θ for all d values of l given these learning parameters

A matrix form would be desired, some equation obtained after rewriting the cost function using matrices

$$J(\theta) = \sum_{i=1}^{n} [h_\theta(x^{(i)}) - y^{(i)}]^2 + \lambda \sum_{j=1}^{d} \theta_j^2$$

$$J(\theta) = \sum_{i=1}^{n} [h_\theta(x^{(i)}) - y^{(i)}]^2 + \lambda \theta^T \theta$$

$$J(\theta) = \left(\theta^T \mathbf{X} - \mathbf{y}\right)^T \left(\theta^T \mathbf{X} - \mathbf{y}\right) + \lambda \theta^T \theta$$

We should be able to find the derivative of the scalar $J(\theta)$ with respect to θ